On conjugate families and Jeffreys priors for von Mises-Fisher distributions

Hornik, Kurt; Grün, Bettina

Published in:
Journal of Statistical Planning and Inference

DOI:
10.1016/j.jspi.2012.11.003

Published: 01/02/2013

Document Version:
Publisher's PDF, also known as Version of record

The following pages contain the PDF file of the publication:

Hornik, Kurt and Grün, Bettina (2013). On conjugate families and Jeffreys priors for von Mises–Fisher distributions. *Journal of Statistical Planning and Inference*, 143(5):992–999. DOI: 10.1016/j.jspi.2012.11.003

The authors regret that the original publication contains an error in the proof of Theorem 4. Theorem 4 remains valid. However, the proof should read:

Proof of Theorem 4. The first assertion is immediate by observing that $\det(I(\theta))$ depends on θ only via its length. To obtain the asymptotic behavior, we can use the asymptotics of $A_d(\kappa)$ and the fact that $A_d'(\kappa) = 1 - A_d(\kappa)(A_d(\kappa) + (d-1)/\kappa)$ (Schou, 1978). Thus, as $\kappa \to \infty$,

$$\begin{aligned}
A_d'(\kappa) &= 1 - \left(1 - \frac{d-1}{2\kappa} + \frac{(d-1)(d-3)}{8\kappa^2} + O(\kappa^{-3})\right) \\
&\qquad \left(1 - \frac{d-1}{2\kappa} + \frac{(d-1)(d-3)}{8\kappa^2} + O(\kappa^{-3}) + \frac{d-1}{\kappa}\right) \\
&= 1 - \left(1 - \frac{d-1}{2\kappa} + \frac{(d-1)(d-3)}{8\kappa^2} + O(\kappa^{-3})\right)\left(1 + \frac{d-1}{2\kappa} + \frac{(d-1)(d-3)}{8\kappa^2} + O(\kappa^{-3})\right) \\
&= 1 - 1 - \frac{d-1}{2\kappa} - \frac{(d-1)(d-3)}{8\kappa^2} + \frac{d-1}{2\kappa} + \frac{(d-1)^2}{4\kappa^2} - \frac{(d-1)(d-3)}{8\kappa^2} + O(\kappa^{-3}) \\
&= \frac{(d-1)^2}{4\kappa^2} - \frac{(d-1)(d-3)}{4\kappa^2} + O(\kappa^{-3}) \\
&= \frac{(d-1)}{4\kappa^2}(d-1-d+3) + O(\kappa^{-3}) \\
&= \frac{(d-1)}{2\kappa^2} + O(\kappa^{-3})
\end{aligned}$$

such that for $\|\theta\| \to \infty$,

$$\det(I(\theta)) \approx \left(\frac{1}{\|\theta\|}\right)^{d-1} \frac{(d-1)}{2\|\theta\|^2} = \frac{\text{const}}{\|\theta\|^{d+1}}.$$

The authors would like to apologize for any inconvenience caused.

Journal of Statistical Planning and Inference 143 (2013) 992–999

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

On conjugate families and Jeffreys priors for von Mises–Fisher distributions

Kurt Hornik [a], Bettina Grün [b,*]

[a] Institute for Statistics and Mathematics, WU Wirtschaftsuniversität Wien, Augasse 2–6, 1090 Vienna, Austria
[b] Department of Applied Statistics, Johannes Kepler University Linz, Altenbergerstraße 69, 4040 Linz, Austria

ARTICLE INFO

Article history:
Received 27 September 2011
Received in revised form
7 August 2012
Accepted 6 November 2012
Available online 23 November 2012

Keywords:
Bayesian inference
Conjugate prior
Jeffreys prior
von Mises–Fisher distribution

ABSTRACT

This paper discusses characteristics of standard conjugate priors and their induced posteriors in Bayesian inference for von Mises–Fisher distributions, using either the canonical natural exponential family or the more commonly employed polar coordinate parameterizations. We analyze when standard conjugate priors as well as posteriors are proper, and investigate the Jeffreys prior for the von Mises–Fisher family. Finally, we characterize the proper distributions in the standard conjugate family of the (matrix-valued) von Mises–Fisher distributions on Stiefel manifolds.

1. Introduction

A random unit length vector in $\mathbb{R}^d$ has a von Mises–Fisher (or Langevin, short: vMF) distribution with parameter $\theta \in \mathbb{R}^d$ if its density with respect to the uniform distribution on the unit hypersphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ is given by

$$f(x|\theta) = e^{\theta' x} / {}_0F_1(;d/2;\|\theta\|^2/4),$$

where, using the rising factorial $(v)_n = \Gamma(v+n)/\Gamma(v)$,

$${}_0F_1(;v;z) = \sum_{n=0}^{\infty} \frac{1}{(v)_n} \frac{z^n}{n!} = \sum_{n=0}^{\infty} \frac{\Gamma(v)}{\Gamma(v+n)} \frac{z^n}{n!}$$

is a generalized hypergeometric series and related to the modified Bessel function of the first kind I_v via

$${}_0F_1(;v+1;\kappa^2/4) = \frac{I_v(\kappa)\Gamma(v+1)}{(\kappa/2)^v}$$

(e.g., Mardia and Jupp, 1999, p. 168).

We note that the vMF distribution is commonly parameterized using polar coordinates, i.e., $\theta = \kappa\mu$, where $\kappa = \|\theta\|$ and $\mu \in \mathbb{S}^{d-1}$ are the concentration and mean direction parameters, respectively (if $\theta \neq 0$, μ is uniquely determined as $\theta/\|\theta\|$). Using θ as the parameter, the family $\mathcal{F}$ of vMF distributions on $\mathbb{S}^{d-1}$ becomes a natural exponential family through the

* Corresponding author.
E-mail addresses: Kurt.Hornik@wu.ac.at (K. Hornik), Bettina.Gruen@jku.at (B. Grün).

0378-3758
http://dx.doi.org/10.1016/j.jspi.2012.11.003

uniform distribution U on $\mathbb{S}^{d-1}$, commonly written as

$$f(x|\theta) = e^{\theta' x - M(\theta)},$$

where in the vMF case, the cumulant transform $M(\theta)$ of U is given by

$$e^{M(\theta)} = \int_{\mathbb{S}^{d-1}} e^{\theta' x} dU(x) = {}_0F_1(;d/2;\|\theta\|^2/4).$$

Bayesian inference for the vMF distribution is first discussed in Mardia and El-Atoum (1976), who give conjugate priors for μ when κ is known, and derive the Jeffreys prior for the polar coordinates (μ, κ) parameterization. Guttorp and Lockhart (1988) introduce a Bayesian approach for finding the direction of a signal based on developing *standard* (e.g., Gutiérrez-Peña and Smith, 1997, Definition 3.1) conjugate priors for the von Mises (vM) distribution (i.e., for $d=2$) using the canonical (θ) parameterization. Damien and Walker (1999) present a full Bayesian analysis of circular data using the vM distribution by employing standard conjugate priors for the polar coordinates (μ, κ) parameterization, and developing a Gibbs sampler for this family of distributions. Nuñez-Antonio and Gutiérrez-Peña (2005) provide a full Bayesian analysis of directional (i.e., $d \geq 2$) data using the vMF distribution, again using standard (μ, κ) conjugate priors and obtaining samples from the posterior using a sampling-importance-resampling method found to outperform Gibbs sampling. Bangert et al. (2010) construct (possibly infinite) mixtures of vMF distributions using standard conjugate priors for the (μ, κ) parameterization and Dirichlet (process) priors for the mixing probabilities.

Interestingly, none of these references explicitly discuss when the employed priors (and respective posteriors) are actually proper, or whether the conjugate families obtained using the θ or (μ, κ) parameterizations are the same. In this paper, we settle these open issues, and also discuss Jeffreys priors for the general $(d \geq 2)$ vMF family (Section 2). We also provide results for (matrix-valued) vMF distributions on Stiefel manifolds (Section 3).

2. Results

2.1. Propriety of priors from the standard conjugate family

In what follows, it will be convenient to write

$$C_d(\kappa) = 1/{}_0F_1(;d/2;\kappa^2/4),$$

so that $e^{-M(\theta)} = C_d(\|\theta\|)$. Let $\theta = \theta(\lambda)$ be a suitable parameterization of θ. For a sample $x_1, \ldots, x_n$ of independent, identically distributed (i.i.d.) observations from the vMF family $\mathcal{F}$, the likelihood function for λ is given by

$$L(\lambda|s,n) = e^{s'\theta(\lambda) - nM(\theta(\lambda))} = C_d(\|\theta(\lambda)\|)^n e^{s'\theta(\lambda)},$$

where $s = x_1 + \cdots + x_n$ is the resultant of the sample. Following Gutiérrez-Peña and Smith (1997, Definition 3.1), the *standard conjugate family* for $\mathcal{F}$ relative to λ, denoted by $\mathcal{C}_\lambda(\mathcal{F})$, has densities

$$\pi(\lambda|s,v) \propto L(\lambda|s,v).$$

Using such a prior with parameters s_0 and v_0 will result in a posterior with parameters $s_n = s_0 + x_1 + \cdots + x_n$ and $v_n = v_0 + n$. As clearly

$$\frac{s_n}{v_n} = \frac{v_0}{v_0+n}\frac{s_0}{v_0} + \frac{n}{v_0+n}\bar{x},$$

v_0 can be interpreted as the prior sample size, and s_n/v_n as a sample size weighted average of the "prior mean" s_0/v_0 and the sample mean $\bar{x}$.

The standard conjugate family $\mathcal{C}_\theta(\mathcal{F})$ relative to the canonical parameter θ has several important properties, in particular the linear relationship between the posterior mean and the sample mean (Diaconis and Ylvisaker, 1979).

We note that the densities $\pi(\lambda|s,v)$ are usually taken relative to the Lebesgue measure, which does not quite fit the needs of the commonly used polar coordinates (μ, κ) parameterization of the vMF family $\mathcal{F}$. Let us generally write η for the reference measure employed. Previous work using the (μ, κ) parameterization seem to take η as the product of the Lebesgue measure on $[0,\infty)$ (for κ) and the uniform distribution U on $\mathbb{S}^{d-1}$ (for μ), i.e., $d\eta \propto d\kappa\, dU(\mu)$. As for $\theta = \kappa\mu$ we have $d\theta = a_d \kappa^{d-1}\, d\kappa\, dU(\mu)$ (where a_d is the area of the unit hypersphere). The latter may be more natural as reference measure, turning the standard conjugate family relative to the polar coordinates (μ, κ) parameterization into the (obvious generalization) of what Gutiérrez-Peña and Smith (1997) call the DY-conjugate family for $\mathcal{F}$ relative to the parameterization.

Let $\mathcal{H}_{\lambda;\eta}$ denote the set of all hyperparameters s and v for which $\pi(\lambda|s,v)$ is a proper distribution on the employed parameter space Λ (using η as reference measure), i.e.,

$$\mathcal{H}_{\lambda;\eta} = \left\{(s,v) : \int_\Lambda C_d(\|\theta(\lambda)\|)^v e^{s'\theta(\lambda)} d\eta(\lambda) < \infty\right\},$$

and let

$$J(\alpha,\beta,\nu) = \int_0^\infty \frac{\kappa^\alpha C_d(\kappa)^\nu}{C_d(\beta\kappa)} d\kappa.$$

We have the following results.

Theorem 1. *For the canonical parameterization θ of the vMF family and the Lebesgue measure as reference measure η,*

$$\mathcal{H}_{\theta;\eta} = \{(s,\nu) : \|s\| < \nu\},$$

and the normalizing constant is the inverse of $a_d J(d-1,\|s\|,\nu)$.

In the following a parameter α is introduced which allows to cover both cases of reference measures when using the (μ, κ) parameterization: the Lebesgue measure (leading to $\alpha = d-1$) and the product of the Lebesgue measure on $[0,\infty)$ and the uniform distribution U on $\mathbb{S}^{d-1}$ which is employed in previous work (leading to $\alpha = 0$). Other choices for α lead to additional possible reference measures.

Theorem 2. *For the polar coordinates (μ, κ) parameterization of the vMF family and the reference measure $d\eta = \kappa^\alpha d\kappa\, dU(\mu)$ with $\alpha \geq 0$,*

$$\mathcal{H}_{\kappa,\mu;\eta} = \{(s,\nu) : \|s\| < \nu \quad \text{or } \|s\| = \nu < 1-2(\alpha+1)/(d-1)\},$$

and the normalizing constant is the inverse of $J(\alpha,\|s\|,\nu)$.

Note that if $d \geq 2$, $1-2(\alpha+1)/(d-1) > 0$ is equivalent to $(d-1)/2 > \alpha+1$ or $(d-3)/2 > \alpha$, which if $\alpha \geq 0$ is only possible if $d \geq 4$. Thus, the set $\{(s,\nu) : \|s\| = \nu < 1-2(\alpha+1)/(d-1)\}$ is non-empty only if $d \geq 4$ and $\alpha < (d-3)/2$, and clearly can only contain points for which $\|s\| = \nu < 1$.

For the proof, we use the following result.

Lemma 1. *If α and β are nonnegative, $J(\alpha,\beta,\nu) < \infty$ if and only if $\beta < \nu$ or $0 \leq \beta = \nu < 1-2(\alpha+1)/(d-1)$.*

Proof of Lemma 1. Using the asymptotic approximation $I_\nu(\kappa) \approx e^\kappa/\sqrt{2\pi\kappa}$ for $\kappa \to \infty$ and ν fixed (e.g., Abramowitz and Stegun, 1972, http://dlmf.nist.gov/10.40), we have

$$C_d(\kappa) \propto \kappa^{d/2-1}/I_{d/2-1}(\kappa) \approx \sqrt{2\pi}\kappa^{(d-1)/2}e^{-\kappa}.$$

Hence, for large κ, the integrand in J is "approximately proportional" to

$$\frac{\kappa^\alpha \kappa^{\nu(d-1)/2} e^{-\nu\kappa}}{(\beta\kappa)^{(d-1)/2} e^{-\beta\kappa}} \propto \kappa^{\alpha+(\nu-1)(d-1)/2} e^{-(\nu-\beta)\kappa}.$$

Thus, the integral diverges if $\nu < \beta$, and converges if $\nu > \beta$. If $\nu = \beta$, convergence requires $-1 > \alpha+(\nu-1)(d-1)/2$, or equivalently, $\nu-1 < -2(\alpha+1)/(d-1)$ as asserted. □

Proof of Theorem 1. Transforming to polar coordinates $\theta = \kappa\mu$, we obtain

$$\begin{aligned}\int_{\mathbb{R}^d} C_d(\|\theta\|)^\nu e^{s'\theta} d\theta &= a_d \int_0^\infty C_d(\kappa)^\nu \left(\int_{\mathbb{S}^{d-1}} e^{\kappa s'\mu} dU(\mu)\right) \kappa^{d-1}\, d\kappa \\ &= a_d \int_0^\infty \kappa^{d-1} C_d(\kappa)^\nu \frac{1}{C_d(\kappa\|s\|)} d\kappa \\ &= a_d J(d-1,\|s\|,\nu),\end{aligned}$$

interchanging the order of integration being justified by nonnegativity of the integrand. The assertion now follows from Lemma 1. □

Proof of Theorem 2. If $d\eta = \kappa^\alpha\, d\kappa\, dU(\mu)$, we have

$$\int_0^\infty \int_{\mathbb{S}^{d-1}} C_d(\kappa)^\nu e^{\kappa s'\mu}\, \kappa^\alpha\, d\kappa\, dU(\mu) = \int_0^\infty \kappa^\alpha \frac{C_d(\kappa)^\nu}{C_d(\kappa\|s\|)}\, d\kappa = J(\alpha,\|s\|,\nu),$$

whence the theorem follows by again using Lemma 1. □

We see that for the canonical parameterization, the hyperparameters giving proper distributions are the ones for which $\nu > 0$ and the "prior mean" s/ν lies in the interior of (the convex hull of) the unit hypersphere $\mathbb{S}^{d-1}$. This is not a coincidence: in fact, one can alternatively establish Theorem 1 (and equivalently, Theorem 2 for $\alpha = d-1$) without explicit convergence computations using the general results of Diaconis and Ylvisaker (1979), see also Gutiérrez-Peña and Smith (1997, Theorem 3.1). Let μ be a probability measure on (the Borel sets of) $\mathbb{R}^d$ with bounded support $\mathcal{S}$ and consider the natural exponential family through μ with density $f(x|\theta) = e^{\theta' x - M(\theta)}$, where $e^{M(\theta)} = \int e^{\theta' x}\, d\mu(x)$, and the standard conjugate family with densities $\pi(\theta|s,\nu) = e^{s'\theta - \nu M(\theta)}$ (with respect to the Lebesgue measure). As $\mathcal{S}$ is bounded and μ is finite, $\Theta = \{\theta : M(\theta) < \infty\} = \mathbb{R}^d$. Let $\mathcal{X}$ be the interior of the convex hull of $\mathcal{S}$. Then by Theorem 1 of Diaconis and Ylvisaker (1979), if $\mathcal{X}$ is nonempty (and hence "the observation set is genuinely d-dimensional" Diaconis and Ylvisaker, 1979, p. 271) $\pi(\theta|s,\nu)$

is proper if and only if $v>0$ and $s/v \in \mathcal{X}$. (The reference actually uses vs where we use s.) In the vMF case, $\mathcal{S}=\mathbb{S}^{d-1}$, with convex hull the closed unit ball, and interior the open unit ball $\mathcal{X}=\{x\in\mathbb{R}^d : \|x\|<1\}$. Hence, $\pi(\theta|s,v)$ is proper if and only if $v>0$ and $\|s/v\|<1$, or equivalently, $\|s\|<v$, again establishing Theorem 1.

We also note that for $\alpha \le d-1$ and $\|s\|<v$, Theorem 1 implies that

$$a_d \int_1^\infty \int_{\mathbb{S}^{d-1}} C_d(\kappa)^v e^{\kappa s'\mu}\kappa^\alpha \, d\kappa \, dU(\mu) \le a_d \int_1^\infty \int_{\mathbb{S}^{d-1}} C_d(\kappa)^v e^{\kappa s'\mu}\kappa^{d-1} \, d\kappa \, dU(\mu) = \int_{\|\theta\|\ge 1} C_d(\|\theta\|)^v e^{s'\theta} \, d\theta < \infty,$$

so that s and v give a proper conjugate distribution for the polar coordinates (μ, κ) parameterization with reference measure $d\eta = \kappa^\alpha \, d\kappa \, dU(\mu)$. However, neither results for the case $\alpha > d-1$ nor necessity of the condition $\|s\|<v$ can be established using the general framework (and in fact, Theorem 2 shows that the condition is not necessary if $d\ge 4$ and $0\le\alpha<(d-3)/2$).

2.2. Propriety of posteriors from improper standard conjugate priors

Quite interestingly, canonical priors employed in the literature are improper if a vague prior is intended (see for example Nuñez-Antonio and Gutiérrez-Peña, 2005, who use $v=0$ and $s=0$). However, we note that if $\|s\|=v>0$ and $x_1,\ldots,x_n \in \mathbb{S}^{d-1}$, then $\|s+x_1+\cdots+x_n\| \le \|s\|+\|x_1\|+\cdots+\|x_n\| \le n+v$ with equality if and only if $x_1=\cdots=x_n=s/v$ which is a zero set for samples obtained from the vMF with fixed parameter θ. Hence intuitively, we expect that improper standard conjugate priors with $\|s\|=v>0$ "almost always" yield proper posteriors. For the case $\|s\|=v=0$ (as in the examples) we have $\|x_1+\cdots+x_n\| \le \|x_1\|+\cdots+\|x_n\| \le n$ with equality if and only if $x_1=\cdots=x_n$ which is a zero set for samples obtained from the vMF with fixed parameter θ and $n\ge 2$.

This can be formalized as follows. Let π be the density (with respect to η) of a σ-finite measure on Λ and define $\mu_\pi^{(n)}$ on $\Xi^{(n)}=\mathbb{S}^{d-1}\times\cdots\times\mathbb{S}^{d-1}$ (n times), the space of all $\mathbb{S}^{d-1}$ valued samples of size n, via

$$\begin{aligned}\mu_\pi^{(n)}(A) &= \int_\Lambda \int_A f(x_1|\theta(\lambda))\cdots f(x_n|\theta(\lambda)) dU(x_1)\cdots dU(x_n)\pi(\lambda)\, d\eta(\lambda)\\ &= \int_A \int_\Lambda f(x_1|\theta(\lambda))\cdots f(x_n|\theta(\lambda))\pi(\lambda)\, d\eta(\lambda)\, dU(x_1)\cdots dU(x_n).\end{aligned}$$

Writing $g_\pi(x_1,\ldots,x_n)=\int_\Lambda f(x_1|\theta(\lambda))\cdots f(x_n|\theta(\lambda))\pi(\lambda)\, d\eta(\lambda)$,

$$\mu_\pi^{(n)}(A) = \int_A g_\pi(x_1,\ldots,x_n) dU(x_1)\cdots dU(x_n),$$

i.e., $\mu_\pi^{(n)}$ is a generalized "mixture" of the distribution of i.i.d. samples of size n from the vMF family. Let

$$A_n(s,v)=\{(x_1,\ldots,x_n) : \|s+x_1+\cdots+x_n\| \ge (n+v)\}.$$

Theorem 3. *If $\|s\|=v>0$ ($s=v=0$), then $\mu_\pi^{(n)}(A_n(s,v))=0$ for all $n\ge 1$ (respectively, ≥ 2) and arbitrary π.*

Proof of Theorem 3. Let $v>0$. From the above, $(x_1,\ldots,x_n)\in A_n(s,v)$ if and only if $x_1=\cdots=x_n=s/v$. Clearly, for i.i.d. random variables $(X_1,\ldots,X_n)$ from the vMF distribution with parameter θ, $\mathbb{P}((X_1,\ldots,X_n)\in A_n(s,v)|\theta) = \int_{A_n(s,v)} f(x_1|\theta)\cdots f(x_n|\theta)\, dU(x_1)\cdots dU(x_n)=0$ and hence, $\mu_\pi^{(n)}(A_n(s,v))=\int_\Lambda 0\cdot\pi(\lambda)\, d\eta(\lambda)=0$. As $A_n(0,0)$ consists of all $(x_1,\ldots,x_n)$ for which $x_1=\cdots=x_n$, the assertion for the second case follows along the lines of the first case. □

If we use the canonical parameterization and $\pi(\theta)=e^{\theta' s - vM(\theta)}$, then by the above,

$$g_\pi(x_1,\ldots,x_n) = \int_\Theta e^{\theta'(x_1+\cdots+x_n)-nM(\theta)} e^{\theta' s - vM(\theta)}\, d\theta = a_d J(d-1, \|s+x_1+\cdots+x_n\|, n+v)$$

is infinite if and only if $(x_1,\ldots,x_n)\in A_n(s,v)$. If $\|s\|=v>0$, this is a zero set under the product of uniforms on $\Xi^{(n)}$, and hence (again) $\mu_\pi^{(n)}(A_n(s,v))=\int_A g_\pi(x_1,\ldots,x_n) dU(x_1)\cdots dU(x_n)=0$. On the other hand, if $\|s\|>v$, then clearly $\mu_\pi^{(n)}(A_n(s,v))=\infty$: in this sense, it is always possible to obtain improper posteriors when employing an improper standard conjugate prior with $\|s\|>v$ (we notice however that such priors are admittedly "strange", as the corresponding prior sample means s/v are outside the unit ball and hence "impossible").

If X has a vMF distribution with parameter θ, the theory of regular exponential models (cf., e.g., Mardia and Jupp, 1999, pp. 32–33) implies that $E_\theta(X)=\partial M(\theta)/\partial\theta'$ so that

$$E_\theta(X) = -\frac{d\log(C_d(\|\theta\|))}{d\|\theta\|}\frac{\partial\|\theta\|}{\partial\theta'} = -\frac{C_d'(\|\theta\|)}{C_d(\|\theta\|)}\frac{\theta}{\|\theta\|} = \frac{A_d(\|\theta\|)}{\|\theta\|}\theta,$$

where one can show that the logarithmic derivative A_d of $1/C_d$ satisfies (Schou, 1978)

$$A_d(\kappa) = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)},$$

that

$$A_d(\kappa) = \frac{1}{d}\kappa - \frac{1}{d^2(d+2)}\kappa^3 + \mathrm{O}(\kappa^5),$$

as $\kappa \to 0$ (so that $A_d(\kappa)/\kappa$ is in fact C^∞ provided we take its value at zero to be $1/d$), and that

$$A_d(\kappa) = 1 - \frac{(d-1)}{2}\frac{1}{\kappa} + \frac{(d-1)(d-3)}{8}\frac{1}{\kappa^2} + \mathrm{O}(\kappa^{-3})$$

as $\kappa \to \infty$. Hence, for i.i.d. random variables $X_1, \ldots, X_n$ from the vMF distribution with parameter θ, $\overline{X}_n \to E_\theta(X) = A_d(\|\theta\|)\theta/\|\theta\|$ which is less than one in length, and hence

$$\frac{\|S_n\|}{v_n} = \frac{\|s + X_1 + \cdots + X_n\|}{n+v} \to \|E_\theta(X)\| < 1,$$

with probability one as $n \to \infty$. Thus, if we write

$$p_n(\theta|s,v) = P_\theta((X_1, \ldots, X_n) \in A_n(s,v)),$$

$p_n(\theta|s,v) \to 0$ as $n \to \infty$ for all θ, and using continuity arguments one can easily see that this convergence is uniform on compact subsets of $\Theta = \mathbb{R}^d$. On the other hand, $\|E_\theta(X)\| \to 1$ for $\|\theta\| \to \infty$, so the convergence cannot be uniform over Θ. It would be very interesting to find the rate at which $\sup_{\|\theta\| \le \kappa} p_n(\theta|s,v)$ tends to zero, which would then allow one to characterize the improper prior densities π for which $\mu_\pi^{(n)}(A_n(s,v)) = \int_\Theta p_n(\theta|s,v)\pi(\theta)\, d\theta \to 0$ as $n \to \infty$.

2.3. Jeffreys prior

A commonly suggested non-informative prior is the Jeffreys prior (Jeffreys, 1961), defined as the square root of the determinant of the Fisher information matrix relative to the parameterization employed. When using the canonical parameter, again by the theory of regular exponential models, $I(\theta) = \mathrm{var}_\theta(X) = \partial^2 M(\theta)/\partial\theta\partial\theta'$ so that

$$\begin{aligned} I(\theta) &= A_d'(\|\theta\|)\frac{\theta}{\|\theta\|}\frac{\theta'}{\|\theta\|} + A_d(\|\theta\|)\left(\frac{1}{\|\theta\|}I_d - \theta\frac{1}{\|\theta\|^2}\frac{\theta'}{\|\theta\|}\right) \\ &= A_d'(\|\theta\|)\frac{\theta}{\|\theta\|}\frac{\theta'}{\|\theta\|} + \frac{A_d(\|\theta\|)}{\|\theta\|}\left(I_d - \frac{\theta}{\|\theta\|}\frac{\theta'}{\|\theta\|}\right) \\ &= \frac{A_d(\|\theta\|)}{\|\theta\|}I_d + \left(A_d'(\|\theta\|) - \frac{A_d(\|\theta\|)}{\|\theta\|}\right)\frac{\theta}{\|\theta\|}\frac{\theta'}{\|\theta\|}, \end{aligned}$$

where I_d denotes the d-dimensional identity matrix. By a well known result from linear algebra, $\det(I_d + \beta\nu\nu') = 1 + \beta\nu'\nu$ and thus $\det(\gamma I_d + \beta\nu\nu') = \gamma^d(1 + \beta\nu'\nu/\gamma)$ and in particular if $\|\nu\| = 1$, $\det(\gamma I_d + \beta\nu\nu') = \gamma^d(1 + \beta/\gamma) = \gamma^{d-1}(\beta + \gamma)$, so that

$$\det(I(\theta)) = \left(\frac{A_d(\|\theta\|)}{\|\theta\|}\right)^{d-1} A_d'(\|\theta\|),$$

generalizing the result obtained in Guttorp and Lockhart (1988) for the case $d = 2$ (the sign in the reference is not correct).

Theorem 4. *The Jeffreys prior $\pi(\theta) \propto \sqrt{\det(I(\theta))}$ for the canonical parameter θ only depends on $\|\theta\|$ and behaves like $1/\|\theta\|^{(d+1)/2}$ for $\theta \to \infty$.*

Proof of Theorem 4. The first assertion is immediate by observing that $\det(I(\theta))$ depends on θ only via its length. To obtain the asymptotic behavior, we can use the asymptotics of $A_d(\kappa)$ and the fact that $A_d'(\kappa) = 1 - A_d(\kappa)(A_d(\kappa) + (d-1)/\kappa)$ (Schou, 1978). Thus, as $\kappa \to \infty$,

$$A_d'(\kappa) \approx 1 - \left(1 - \frac{d-1}{2\kappa}\right)\left(1 - \frac{d-1}{2\kappa} + \frac{d-1}{\kappa}\right) = 1 - \left(1 - \frac{d-1}{2\kappa}\right)\left(1 + \frac{d-1}{2\kappa}\right) = \frac{(d-1)^2}{4\kappa^2},$$

such that for $\|\theta\| \to \infty$,

$$\det(I(\theta)) \approx \left(\frac{1}{\|\theta\|}\right)^{d-1}\frac{(d-1)^2}{4\|\theta\|^2} = \frac{\mathrm{const}}{\|\theta\|^{d+1}}. \quad \square$$

We note that the Jeffreys prior "looks different" from the densities employed in the standard conjugate family $\mathcal{C}_\theta(\mathcal{F})$ relative to the canonical parameter. Following Gutiérrez-Peña and Smith (1997), one can rigorously establish that it is not contained in this family by verifying that the skewness vector $\partial \log(\det(I(\theta)))/\partial\theta'$ of the vMF family is not linear in its mean parameter, which is straightforward from the above expression for $\det(I(\theta))$.

The Jeffreys prior with respect to the canonical parameterization is given by

$$\pi(\theta) \propto \sqrt{\left(\frac{A_d(\|\theta\|)}{\|\theta\|}\right)^{d-1} A_d'(\|\theta\|)} = \left(\frac{A_d(\|\theta\|)}{\|\theta\|}\right)^{(d-1)/2}\left(1 - A_d(\|\theta\|)\left(A_d(\|\theta\|) + \frac{(d-1)}{\|\theta\|}\right)\right)^{1/2}.$$

If the polar coordinates (μ, κ) parameterization is employed two alternative parameterizations are possible for the mean direction parameter μ in order to obtain an unrestricted set of parameters. The (μ_0', κ) parameterization with $\mu_0 = (\mu_1, \ldots, \mu_{d-1})'$ consists of the first $d-1$ dimensions of the mean direction parameter μ and the (m', κ) parameterization uses the spherical polar coordinates for μ with $m = (\phi_1, \ldots, \phi_{d-1})'$. If these parameterizations are used the Jeffreys prior derived for the canonical parameterization needs to be multiplied with the Jacobians which are given by

$$\kappa^{d-1}/\mu_d,$$

for the (μ_0', κ) parameterization and

$$\kappa^{d-1} \prod_{j=1}^{d-2} \sin(\phi_j)^{d-1-j},$$

for the (m', κ) parameterization. Note that the latter is also given in Mardia and El-Atoum (1976), which should have $\kappa - (d-1)A_d(\kappa) - \kappa A_d^2(\kappa)$ instead of $\kappa - A_d(\kappa) - \kappa A_d^2(\kappa)$ and needs $d-1-j$ instead of $d-2$ in the exponents of the sinuses.

Clearly, the Jeffreys prior is not proper. The following shows that "almost all" posteriors obtained from it (and in fact, from arbitrary possibly improper priors which increase at most polynomially in $\|\theta\|$) are proper for samples of size $n \geq 2$.

Theorem 5. *Consider the canonical parameterization of the vMF family with the Lebesgue reference measure. Let $\pi(\theta)$ be $O(\|\theta\|^\gamma)$ for some finite $\gamma \geq 0$ as $\|\theta\| \to \infty$ and $B_n(\pi) = \{(x_1, \ldots, x_n) : \int_\Theta f(x_1|\theta) \cdots f(x_n|\theta)\pi(\theta)\, d\theta < \infty\}$. Then for all $n \geq 2$, $\mu_\pi^{(n)}(B_n(\pi)^c) = 0$.*

Proof of Theorem 5. Writing $s = x_1 + \cdots + x_n$, we have

$$\int_{\|\theta\| \geq 1} f(x_1|\theta) \cdots f(x_n|\theta)\pi(\theta)\, d\theta = \int_{\|\theta\| \geq 1} e^{\theta'(x_1 + \cdots + x_n) - nM(\theta)} \pi(\theta)\, d\theta$$
$$\leq \text{const} \int_1^\infty \int_{\mathbb{S}^{d-1}} e^{\kappa s'\mu} C_d(\kappa)^n \kappa^\gamma\, \kappa^{d-1} dU(\mu)\, d\kappa = \text{const} \int_1^\infty \kappa^{\gamma+d-1} \frac{C_d(\kappa)^n}{C_d(\kappa\|s\|)}\, d\kappa.$$

By Lemma 1, this is finite provided that $\|s\| < n$. Hence, $B_n(\pi)^c \subseteq A_n(0,0)$ which is a zero set under the product of uniforms on $\Xi^{(n)}$ provided that $n \geq 2$. □

2.4. Propriety of prior and posterior distributions in applications

Guttorp and Lockhart (1988) perform a full Bayesian analysis employing the canonical parameterization for 2-dimensional data. Rather than using the standard conjugate prior for θ, they use a flat prior on μ and the conjugate prior they derived for κ with μ known. Note that for the conjugate prior for κ to be proper, the same conditions on $\|s\|$ and v need to be satisfied as for the conjugate prior for θ. In order to parameterize the κ prior only the length of s and v need to be specified. For their application, Guttorp and Lockhart (1988) use three different prior distributions for κ: a data-based, a high precision and a low precision prior. In all three cases the priors for κ are proper because $\|s\| < v$. This is also clear by construction: the parameters of the priors are determined by specifying constraints for the moments of the prior distribution or quantities derived from the moments.

Damien and Walker (1999) employ the polar coordinates (μ, κ) parameterization with the Lebesque measure as reference measure, i.e., $\alpha = 0$. They use two different prior distributions in their two numerical examples. In the first example, they set all prior parameters to zero. This is the same prior Nuñez-Antonio and Gutiérrez-Peña (2005) use in their examples and refer to as vague prior. Given the updates for the posterior parameters as well as the interpretation of v as the prior sample size, this seems to be an obvious choice. This prior is improper, but as shown in Section 2.2, posteriors will be proper almost surely for samples of size $n \geq 2$. In their second example, Damien and Walker (1999) use a flat prior for μ and a conjugate prior for κ with μ known. Referring to the low precision prior in Guttorp and Lockhart (1988), $\|s\| = v = 5$ are employed as parameters. The low precision prior in Guttorp and Lockhart (1988) actually is equal to $\|s\| = 9.8824$ and $v = 10$, while the values for the high precision prior are $\|s\| = 4.99893$ and $v = 5$. Interpreting v as the prior sample size, larger values of v imply more informative priors. However, the precision induced by the prior will depend on the average length given by $\|s\|/v$. The closer this value is to 1 the higher is the precision induced by the prior. Using an approximation for large κ, Guttorp and Lockhart (1988) derived that the prior mean and variance of the precision parameter κ depend on v as well as the difference $v - \|s\|$. By rounding $\|s\|$ to the same value as v, Damien and Walker (1999) use an improper prior and the interpretation as a low precision prior, as induced by the prior mean and standard deviation, obviously is lost. As shown in Section 2.2, the posterior from this prior is almost surely proper for samples of size $n \geq 1$.

Bangert et al. (2010) use conditionally conjugate priors for μ and κ. The prior for μ is a von Mises–Fisher distribution with precision parameter equal to 0.1 and mean parameter equal to the mean direction of the data. For κ they use the conjugate prior for known μ. Setting the prior parameters equal to $v = 5$ and $\|s\| = 4.7$, they employ a proper prior.

To sum up, these previous applications indicate that the non-informative but improper prior with $\|s\| = v = 0$ seems to be an obvious choice if no prior information is available. This seems to be unproblematic because the posteriors will be almost surely proper for sample sizes $n \geq 2$ by Theorem 3. If prior information on the precision parameter κ is to be included, a flat prior for μ is employed and the conjugate prior with known μ for κ. Another possibility, when using Gibbs

sampling for estimation, is to employ conditionally conjugate priors. In general the parameters of the prior of κ are chosen to reflect the prior information available for the moments of κ, which leads to proper priors.

3. Extensions

The vMF family on $\mathbb{S}^{d-1}$ can straightforwardly be generalized to the vMF family on the Stiefel manifold $V_k(\mathbb{R}^d)$, the set of orthogonal k-frames in $\mathbb{R}^d$, or equivalently,

$$V_k(\mathbb{R}^d) = \{X \in \mathbb{R}^{d\times k} : X'X = I_k\},$$

so that $V_1(\mathbb{R}^d)$ corresponds to $\mathbb{S}^{d-1}$. The vMF family on $V_k(\mathbb{R}^d)$ has densities

$$f(X|A) = e^{\mathrm{tr}(A'X)-M(A)},$$

with respect to the uniform distribution U on $V_k(\mathbb{R}^d)$, where

$$e^{M(A)} = \int_{V_k(\mathbb{R}^d)} e^{\mathrm{tr}(A'X)} dU(X) = {}_0F_1(;d/2;A'A/4)$$

is a generalized hypergeometric function with matrix argument (e.g., Mardia and Jupp, 1999, p. 289). This family of distributions is useful as a probability distribution over orthonormal matrices and for example Hoff (2009) indicates that it arises as a posterior distribution for the orthonormal matrices in factor analysis when uniform priors are used. For a further discussion of this family of distributions in relation to orientation statistics see Downs (1972) and Khatri and Mardia (1977).

Clearly, $X \mapsto \mathrm{vec}(X)$ defines a one-to-one correspondence between $\mathbb{R}^{d\times k}$ and $\mathbb{R}^{dk}$ with $\mathrm{tr}(X'A) = \mathrm{vec}(X)'\mathrm{vec}(A)$. The standard conjugate family for the vMF family on $V_k(\mathbb{R}^d)$ (relative to the canonical parameter) is thus given by the family of densities

$$p(A|S,\nu) \propto e^{\mathrm{tr}(S'A)-\nu M(A)}.$$

Let $\|S\|_2$ denote the spectral norm (matrix 2-norm, the largest singular value) of S.

Theorem 6. *The distributions in the standard conjugate family of the vMF family on the Stiefel manifold $V_k(\mathbb{R}^d)$ are proper if and only if $\|S\|_2 < \nu$.*

Proof of Theorem 6. The support of U is $\mathcal{S} = V_k(\mathbb{R}^d)$, the convex hull of which is the closed unit ball in the spectral norm (e.g., Journée et al., 2010 or Gallivan and Absil, 2010), and hence has non-empty interior

$$\mathcal{X} = \{X \in \mathbb{R}^{d\times k} : \|X\|_2 < 1\} = \{X \in \mathbb{R}^{d\times k} : X'X \prec I_k\}.$$

Using Theorem 1 of Diaconis and Ylvisaker (1979), the standard conjugate distributions are proper if and only if $\nu > 0$ and $S/\nu \in \mathcal{X}$, or equivalently, if and only if $\|S\|_2 < \nu$. □

The matrix vMF distributions are typically parameterized using the canonical parameter A. Alternatively, the analogue to the polar coordinates (μ, κ) parameterization in the vector case is using the (right) polar decomposition of $A = MK$, where the *polar part* (or *orientation*) M is in the Stiefel manifold $V_k(\mathbb{R}^d)$ and the *elliptical part* (or *concentration*) K is a symmetric, non-negative definite matrix (e.g., Mardia and Jupp, 1999, p. 286).

If A has full rank, K is the unique symmetric matrix root of $A'A$, and (e.g., Cadet, 1996, adjusting for the different normalizations employed) $dA = c(K)dK dU(M)$, where $c(K) = a_{d,k}\det(K)^{d-k}\prod_{i<j}(\lambda_i+\lambda_j)$ with $a_{d,k}$ the (generalized) volume of $V_k(\mathbb{R}^d)$ and $\lambda_1,\ldots,\lambda_k > 0$ the eigenvalues of K. Hence,

$$\begin{aligned}\int_{\mathbb{R}^{d\times k}} e^{\mathrm{tr}(S'A)-\nu M(A)}\, dA &= \int_{K\succcurlyeq 0}\int_{V_k(\mathbb{R}^d)} e^{\mathrm{tr}(S'MK)}({}_0F_1(;d/2;K^2/4))^{-\nu} c(K)\, dK\, dU(M)\\ &= \int_{K\succcurlyeq 0} c(K)({}_0F_1(;d/2;K^2/4))^{-\nu}\int_{V_k(\mathbb{R}^d)} e^{\mathrm{tr}(KS'M)} dU(M)\, dK = \int_{K\succcurlyeq 0} c(K)\frac{{}_0F_1(;d/2;KS'SK/4)}{{}_0F_1(;d/2;K^2/4)^{\nu}}\, dK.\end{aligned}$$

Thus, if we consider the standard conjugate family of the matrix vMF family on the Stiefel manifold $V_k(\mathbb{R}^d)$ relative to the polar coordinates parameterization $A = MK$ with elements $\pi(M,K|S,\nu) \propto e^{\mathrm{tr}(S'MK)}/{}_0F_1(;d/2;K^2/4)^{\nu}$, and reference measures of the form $c(K)^{\alpha}\, dK\, dU(M)$, then as discussed in Section 2 for the polar parameterization of the vector vMF distribution, if $0 \le \alpha \le 1$ Theorem 6 implies that distributions in this conjugate family are proper provided that $\|S\|_2 < \nu$. Again, necessity of this condition for such values of α, or the characterization of the hyperparameters giving proper distributions if $\alpha > 1$ cannot be established.

For this, one needs to be able to characterize S and ν (and α) for which

$$J_k(\alpha,S,\nu) = \int_{K\succcurlyeq 0} c(K)^{\alpha}\frac{{}_0F_1(;d/2;KS'SK/4)}{{}_0F_1(;d/2;K^2/4)^{\nu}}\, dK < \infty,$$

which seems quite challenging, requiring suitable "large K" asymptotics for ${}_0F_1(;d/2;K^2/4)$ and ${}_0F_1(;d/2;KS'SK/4)$. We note that Butler and Wood (2003) give Laplace approximations for ${}_0F_1$ (and corresponding Bessel functions) of matrix arguments (but do not formally establish validity as an asymptotic approximation). Muirhead (1978, p. 22) gives an

asymptotic approximation for ${}_0F_1(;d/2;A'A/4)$ for the case where all singular values of A are large. For the above, a generalization to the case where *some* singular values are large is needed. We leave this for future research.

Acknowledgments

This research was funded by the Austrian Science Fund (FWF): V170-N18.

References

Abramowitz, M., Stegun, I.A., 1972. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, New York.
Bangert, M., Hennig, P., Oelfke, U., 2010. Using an infinite von Mises–Fisher mixture model to cluster treatment beam directions in external radiation therapy. In: Proceedings of the Ninth International Conference on Machine Learning and Applications (ICMLA), pp. 746–751.
Butler, R.W., Wood, A.T.A., 2003. Laplace approximation for Bessel functions of matrix argument. Journal of Computational and Applied Mathematics 155, 359–382.
Cadet, A., 1996. Polar coordinates in $\mathbf{R}^{np}$; application to the computation of the Wishart and beta laws. Sankhya: The Indian Journal of Statistics Series A 58 (1), 101–114.
Damien, P., Walker, S., 1999. A full Bayesian analysis of circular data using the von Mises distribution. The Canadian Journal of Statistics 27 (2), 291–298.
Diaconis, P., Ylvisaker, D., 1979. Conjugate priors for exponential densities. The Annals of Statistics 7 (2), 269–281.
Downs, T.D., 1972. Orientation statistics. Biometrika 59 (3), 665–676.
Gallivan, K.A., Absil, P.-A., 2010. Note on the Convex Hull of the Stiefel Manifold. Technical Report. FSU 10-06, Department of Mathematics, Florida State University. URL ⟨http://www.math.fsu.edu/∼aluffi/archive/paper386.pdf⟩.
Gutiérrez-Peña E, Smith, A.F.M., 1997. Exponential and Bayesian conjugate families: review and extensions. Test 6, 1–90.
Guttorp, P., Lockhart, R.A., 1988. Finding the location of a signal: a Bayesian analysis. Journal of the American Statistical Association 83 (402), 322–330.
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, 438–456.
Jeffreys, H., 1961. Theory of Probability, 3rd ed. Oxford University Press, Oxford.
Journée, M., Nesterov, Y., Richtárik, P., Sepulchre, R., 2010. Generalized power method for sparse principal component analysis. Journal of Machine Learning Research 11, 517–553.
Khatri, C.G., Mardia, K.V., 1977. The von Mises–Fisher matrix distribution in orientation statistics. Journal of the Royal Statistical Society B 39 (1), 95–106.
Mardia, K.V., El-Atoum, S.A.M., 1976. Inference for the von Mises–Fisher distribution. Biometrika 63 (1), 203–206.
Mardia, K.V., Jupp, P.E., 1999. Directional Statistics. Probability and Statistics. Wiley.
Muirhead, R.J., 1978. Latent roots and matrix variates: a review of some asymptotic results. The Annals of Statistics 2 (1), 5–33.
Nuñez-Antonio, G., Gutiérrez-Peña, E., 2005. A Bayesian analysis of directional data using the von Mises–Fisher distribution. Communications in Statistics—Simulation and Computation 34 (4), 989–999.
Schou, G., 1978. Estimation of the concentration parameter in von Mises–Fisher distributions. Biometrika 65 (1), 369–377.