View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by RERO DOC Digital Library

Stat Comput (2008) 18: 321–331
DOI 10.1007/s11222-008-9060-4

Some computational aspects of the generalized von Mises distribution

Riccardo Gatto

Received: 28 April 2007 / Accepted: 11 March 2008 / Published online: 25 March 2008
© Springer Science+Business Media, LLC 2008

Abstract This article deals with some important computational aspects of the generalized von Mises distribution in relation with parameter estimation, model selection and simulation. The generalized von Mises distribution provides a flexible model for circular data allowing for symmetry, asymmetry, unimodality and bimodality. For this model, we show the equivalence between the trigonometric method of moments and the maximum likelihood estimators, we give their asymptotic distribution, we provide bias-corrected estimators of the entropy, the Akaike information criterion and the measured entropy for model selection, and we implement the ratio-of-uniforms method of simulation.

Keywords Circular distribution · Akaike information criterion · Efficient score · Entropy · Fisher information · Fourier series · Kullback-Leibler information · Maximum likelihood estimator · Mixture distribution · Ratio-of-uniforms method · Trigonometric method of moments estimator

1 Introduction

In various scientific fields observations are directions in two or three dimensions and are referred to as "directional data". Two-dimensional directions are also called "circular data". Besides two-dimensional directions, any periodic phenomenon with a known period leads to circular data. There are many examples of circular data: wind directions, directions of migratory birds, orientation of rock cores, daily arrival times, etc. The circular distribution of a random angle θ in radian measure is defined by $F(x) = \mathrm{P}[0 < \theta \leq x]$, if $x \in [0, 2\pi)$, and $F(x + 2\pi) - F(x) = 1$, if $x \in \mathrm{R}$. Thus $F(0) = 0$, $F(2\pi) = 1$ and for $x_1 \leq x_2 \leq x_1 + 2\pi$, $\mathrm{P}[x_1 < \theta \leq x_2] = F(x_2) - F(x_1) = \int_{x_1}^{x_2} dF(x)$. When F is absolutely continuous, then there exists a circular density f satisfying $F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$. With circular data, it is common practice to use the von Mises (vM) distribution (also called circular normal) given by (2) below, although it does not provide sufficient flexibility for many scientific problems: it is always circularly symmetric, unimodal and with density dropping exponentially on either side from the center.

To overcome this problem while maintaining the important theoretical properties of the vM distribution, Gatto and Jammalamadaka (2007) analyzed the generalized von Mises (GvM) distribution which originates from Maksimov (1967). The generalized von Mises density of order k (GvM_k) is defined as

$$f(\omega \mid \mu_1, \ldots, \mu_k, \kappa_1, \ldots, \kappa_k) = \frac{1}{2\pi G_0^{(k)}(\delta_1, \ldots, \delta_{k-1}, \kappa_1, \ldots, \kappa_k)} \times \exp\left\{\sum_{j=1}^{k} \kappa_j \cos j(\omega - \mu_j)\right\}, \quad (1)$$

where $\kappa_1, \ldots, \kappa_k \geq 0$, $\mu_1 \in [0, 2\pi)$, $\mu_2 \in [0, \pi)$, $\ldots$, $\mu_k \in [0, 2\pi/k)$, and where

$$G_0^{(k)}(\delta_1, \ldots, \delta_{k-1}, \kappa_1, \ldots, \kappa_k)$$

R. Gatto (✉)
Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, 3012 Bern, Switzerland
e-mail: gatto@stat.unibe.ch
url: http://www.stat.unibe.ch/~gatto

$$= \frac{1}{2\pi}\int_0^{2\pi} \exp\{\kappa_1 \cos\omega + \kappa_2 \cos 2(\omega+\delta_1) + \cdots + \kappa_k \cos k(\omega+\delta_{k-1})\}d\omega$$

is a generalization of the Bessel function I_0 (see below) with $\delta_1 = (\mu_1 - \mu_2) \mathrm{mod}\pi$, $\delta_2 = (\mu_1 - \mu_3) \mathrm{mod}(2\pi/3)$, …, $\delta_{k-1} = (\mu_1 - \mu_k) \mathrm{mod}(2\pi/k)$. We denote as $\theta \sim \mathrm{GvM}_k(\mu_1, \ldots, \mu_k, \kappa_1, \ldots, \kappa_k)$ that the circular random variable θ has the density (1). We focus here on the important practical case of $k = 2$, which gives the closest density amongst circular densities with fixed second trigonometric moment, see (3) below, to the von Mises (vM) density (2), the closeness being in the Kullback-Leibler sense (Gatto 2008, Corollary 2.2). GvM_2 densities allow for asymmetry and bimodality, see Gatto and Jammalamadaka (2007, Figs. 1 and 2) for some typical graphs. Note that bimodality is an essential property for modeling wind directions, where it is typical to observe opposite directions, and the GvM_2 model is important in this context, as illustrated in Sect. 4.

The well-known vM density is obtained by interrupting the summation in the exponent of (1) at $k = 1$, giving

$$f(\omega \mid \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(\omega - \mu)\}, \tag{2}$$

for $\mu \in [0, 2\pi)$ and $\kappa \geq 0$, where $I_r(z) = (2\pi)^{-1}\int_0^{2\pi} \cos r\omega \times \exp\{z\cos\omega\}\, d\omega$, $z \in \mathrm{C}$, is the modified Bessel function I of integer order r (see Abramowitz and Stegun 1972, 9.6.19, p. 376). GvM_2 densities maintain most of the theoretical of vM densities and allow for asymmetry and bimodality. The burden for this is an increased complexity of the normalizing constant $G_0^{(2)}$, which will be simply denoted by G_0 and which can be easily evaluated by (16). Other important circular distributions like the wrapped distributions can in general be expressed as infinite sums only, thus they have another kind of undesired complexity. Flexible circular distributions can also be obtained by finite mixtures of simpler distributions, like mixtures of vM distributions (2). Here again, the flexibility proposed by mixtures does not come gratuitously: these mixtures bring more complicated computational procedures and other inferential complications due to their lack of sufficiency, invariance, etc. Mixture models are also a typical source of non-regular maximum likelihood problems. Mixtures of vM distributions do not share the important theoretical properties inherent to the GvM_k distributions, which can be summarized as follows.

- The GvM_k distributions belong to the canonical exponential class (12) after applying the reparameterization (10). Some of the facts given below are consequences of this.
- Once reparameterized, a GvM_k distribution admits a minimal sufficient and complete statistic and it is given by (11) below.
- For the GvM_2 distribution, we show in Sect. 2.1 that the maximum likelihood estimator (MLE) of the parameters is equivalent to the trigonometric method of moments estimator (TMME). The computation of the MLE is conceptually simpler with the GvM_2 distribution than with a mixture of vM distributions.
- In Gatto (2008) it is shown that GvM_k distributions possess some interesting information theoretic properties. An important result is that, under constraints on the trigonometric moments, the closest circular distribution to any fixed circular distribution has the GvM_k form. This closeness property is useful e.g. in Bayesian statistics, whenever the closest prior distribution to a given distribution must be selected and information on some trigonometric moments is available.
- A practical formula for the entropy of the GvM_2 distribution is available and provided by (18).

Concerning mixtures of vM distributions, we should be aware of the following facts.

- While the likelihood function of any GvM_k distribution is bounded, the likelihood of the mixture of e.g. the $\mathrm{vM}(\mu_1, \kappa_1)$ and the $\mathrm{vM}(\mu_2, \kappa_2)$ distributions is unbounded. To see this, consider e.g. $\kappa_1 \to \infty$, then the likelihood at μ_1 equal to any of the sample values goes to infinity. (For this, note that $I_0(\kappa) \sim (2\pi\kappa)^{-1/2} e^{\kappa}$ as $\kappa \to \infty$, see Abramowitz and Stegun (1972, 9.7.1, p. 377).) A bounded likelihood is required in the proof of consistency of the MLE, see e.g. Cox and Hinkley (1974, p. 289). The overall supremum of the likelihood of a vM mixture does not provide a sensible (i.e. consistent) estimator, although some other local supremum do so. The general problem of unbounded likelihood is reviewed by Cheng and Traylor (1995, Sect. 3), who mention the modified likelihood approach by Cheng and Iles (1987) and the spacings-based approach by Cheng and Amin (1983) and Ranneby (1984) as being the "least subjective" solutions to this problem. Alternative estimators for vM mixtures are also given by Spurr and Koutbeiy (1991).
- With the GvM_k distributions, the likelihood ratio test statistic has a simple form and it is asymptotically chi-squared distributed. The likelihood ratio test for testing the vM distribution against the GvM_2 is described in the last paragraph of Sect. 2.1. On the other side, when testing for e.g. a mixture of two vM distributions against a single vM, the likelihood ratio test is not asymptotically chi-squared distributed, see Hartigan (1985) and Titterington et al. (1985, Sect. 5.4).
- Given a model with unknown parameters $\underline{\eta} = (\eta_1, \ldots, \eta_p)^{\mathrm{T}}$, indeterminacy is the existence of a reparameterization $\underline{\varphi} = \underline{\varphi}(\underline{\eta})$ and of two non-empty disjoint subsets I and J of $\{1, \ldots, p\}$, such that $\varphi_i = 0\ \forall i \in I$ implies that the likelihood is independent of $\varphi_j\ \forall j \in J$, see Cheng and Traylor (1995, p. 14). A consequence of indeterminacy is

the unstable behavior of some parameter estimators. Let us consider the five-parameter mixture of vM densities (2) given by $qf(\omega|\mu,\kappa)+(1-q)f(\omega|\nu,\rho)$ and the reparameterization ν, ρ, $\delta_\mu=(\nu-\mu)\mathrm{mod}(2\pi)$ and $\delta_\kappa=\rho-\kappa$. We can observe the following: if $q=0$ then δ_μ and δ_κ are indeterminate, and if $\delta_\mu=0$ and $\delta_\kappa=0$ then q is indeterminate. Thus estimators may be unstable whenever they approach these critical values.

In the context of this last remark, we can note from (1) that a GvM_k distribution possesses the indeterminacy that if $\kappa_j=0$ then μ_j is indeterminate, $j=1,\ldots,k$. This type of indeterminacy, which appears also in the mixture of vM distributions, is however weaker than the previous indeterminacy involving the mixing parameter q of the vM distributions. Firstly because this last type of indeterminacy can be removed by re-reparameterization, precisely by the Cartesian re-parameterization (10), and secondly because this indeterminacy is only one-directional, i.e. the implication from the null to the indeterminate parameters cannot be reversed, as it happens with the indeterminacy involving the mixing parameter q. This indeterminacy can also lead to a problem in estimation, which is illustrated through a real data example in Sect. 4, where the estimator of κ_1 is almost zero and the empirical Fisher information matrix is almost singular, see (30).

This article deals with some important computational aspects of the GvM_2 distribution: the estimation of the model parameters, the inference for these parameters, the estimation of the entropy, the problem of model selection and the generation of pseudo-random numbers. In Sect. 2 we give the TMME and we show that it is equivalent to the MLE. We then provide asymptotically unbiased estimators of the entropy and formulas for the Akaike information criterion (AIC) and for the measured entropy (ME) of model selection. In Sect. 3, we present two types of acceptance-rejection algorithms for generating GvM_2 pseudo-random numbers. We end the article in Sect. 4 with some numerical illustrations.

2 Estimation and model selection

In Sect. 2.1 we first show the equivalence between the TMME and the MLE under the GvM_2 model and we provide the estimating equations for these estimators. We also give the empirical Fisher information matrix and hence the asymptotic distribution of the estimators. In Sect. 2.2 we give an analytical formula for the entropy of the GvM_2 distribution and two bias-corrected estimators of it when the model parameters are unknown. We then provide formulas for the AIC and the ME of model selection.

2.1 Trigonometric method of moments and maximum likelihood estimators

The TMME is the circular version of the method of moments estimator of linear data. For $r=1,2,\ldots$, the rth trigonometric moment of any circular random variable θ is defined as

$$\varphi_r=\mathrm{E}[e^{\iota r\theta}]=\rho_r e^{\iota\nu_r}=\gamma_r+\iota\sigma_r, \tag{3}$$

where $\rho_r=(\mathrm{E}^2[\cos r\theta]+\mathrm{E}^2[\sin r\theta])^{1/2}$, $\nu_r=\arg\{\mathrm{E}[\cos r\theta],\mathrm{E}[\sin r\theta]\}$, $\gamma_r=\mathrm{E}[\cos r\theta]$ is the rth cosine moment and $\sigma_r=\mathrm{E}[\sin r\theta]$ is the rth sine moment. Suppose $\theta_1,\ldots,\theta_n$ are independent replications of θ whose distribution has p unknown parameters to estimate. For $r=1,\ldots,k$, $k=\lfloor(p+1)/2\rfloor$, $\lfloor x\rfloor$ denoting the largest integer smaller than or equal to x, we equate ρ_r and ν_r to their sample versions $\hat\rho_r=n^{-1}([\sum_{i=1}^n\cos r\theta_i]^2+[\sum_{i=1}^n\sin r\theta_i]^2)^{1/2}$ and $\hat\nu_r=\arg\{\sum_{i=1}^n\cos r\theta_i,\sum_{i=1}^n\sin r\theta_i\}$ respectively, and we solve these equations for the p unknown parameters. Equivalently, we can equate γ_r and σ_r to their sample versions $\hat\gamma_r=n^{-1}\sum_{i=1}^n\cos r\theta_i$ and $\hat\sigma_r=n^{-1}\sum_{i=1}^n\sin r\theta_i$, $r=1,\ldots,k$, and solve for the p unknown parameters. For both variants, the solution with respect to the p unknown parameters yields a TMME based on the k first trigonometric moments. Note that we have a superfluous equation whenever p is odd. This method was suggested by Gatto and Jammalamadaka (2003) for wrapped α-stable distributions.

Now we suppose that $\theta_1,\ldots,\theta_n$ are n independent $\mathrm{GvM}_2(\mu_1,\mu_2,\kappa_1,\kappa_2)$ circular random variables with trigonometric moments $\varphi_r=\rho_r e^{\iota\nu_r}=\gamma_r+\iota\sigma_r$, $r=1,2\ldots$. Let us define

$$\begin{aligned}G_r(\delta,\kappa_1,\kappa_2)&=\frac{1}{2\pi}\int_0^{2\pi}\cos r\omega\exp\{\kappa_1\cos\omega\\&\quad+\kappa_2\cos 2(\omega+\delta)\}d\omega,\\H_r(\delta,\kappa_1,\kappa_2)&=\frac{1}{2\pi}\int_0^{2\pi}\sin r\omega\exp\{\kappa_1\cos\omega\\&\quad+\kappa_2\cos 2(\omega+\delta)\}d\omega,\end{aligned} \tag{4}$$

$$A_r(\delta,\kappa_1,\kappa_2)=\frac{G_r(\delta,\kappa_1,\kappa_2)}{G_0(\delta,\kappa_1,\kappa_2)} \tag{5}$$

and

$$B_r(\delta,\kappa_1,\kappa_2)=\frac{H_r(\delta,\kappa_1,\kappa_2)}{G_0(\delta,\kappa_1,\kappa_2)}, \tag{6}$$

for $r=0,1,\ldots$, where $\delta\in[0,\pi)$ and $\kappa_1,\kappa_2\geq 0$. All these functions can be evaluated with the help of the expansions (16) and (17) below. With this, $\varphi_r=e^{\iota r\mu_1}\{A_r(\delta,\kappa_1,\kappa_2)+\iota B_r(\delta,\kappa_1,\kappa_2)\}$, or, equivalently,

$$\begin{pmatrix}\gamma_r\\\sigma_r\end{pmatrix}=R(r\mu_1)\begin{pmatrix}A_r(\delta,\kappa_1,\kappa_2)\\B_r(\delta,\kappa_1,\kappa_2)\end{pmatrix}, \tag{7}$$

where

$$R(r\mu_1) = \begin{pmatrix} \cos r\mu_1 & -\sin r\mu_1 \\ \sin r\mu_1 & \cos r\mu_1 \end{pmatrix},$$

$r = 1, 2, \ldots$, are rotation matrices and where $\delta = (\mu_1 - \mu_2) \mathrm{mod} \pi$. Hence the TMME of δ, μ_1, κ_1 and κ_2, denoted $\hat{\delta} \in [0, \pi)$ $\hat{\mu}_1 \in [0, 2\pi)$, $\hat{\kappa}_1$ and $\hat{\kappa}_2$, are obtained by replacing γ_r and σ_r by $\hat{\gamma}_r$ and $\hat{\sigma}_r$, $r = 1, 2$, in (7). We define also $\hat{\mu}_2 = (\hat{\mu}_1 - \hat{\delta}) \mathrm{mod} \pi$. If we define the score function

$$\underline{\psi}(\omega; \delta, \mu_1, \kappa_1, \kappa_2) = \begin{pmatrix} \cos\omega - \cos\mu_1 A_1(\delta, \kappa_1, \kappa_2) + \sin\mu_1 B_1(\delta, \kappa_1, \kappa_2) \\ \sin\omega - \cos\mu_1 B_1(\delta, \kappa_1, \kappa_2) - \sin\mu_1 A_1(\delta, \kappa_1, \kappa_2) \\ \cos 2\omega - \cos 2\mu_1 A_2(\delta, \kappa_1, \kappa_2) + \sin 2\mu_1 B_2(\delta, \kappa_1, \kappa_2) \\ \sin 2\omega - \cos 2\mu_1 B_2(\delta, \kappa_1, \kappa_2) - \sin 2\mu_1 A_2(\delta, \kappa_1, \kappa_2) \end{pmatrix}, \tag{8}$$

then it follows from (7) that $\mathrm{E}[\underline{\psi}(\theta_1; \delta, \mu_1, \kappa_1, \kappa_2)] = 0$, the expectation being taken under $\delta, \mu_1, \kappa_1, \kappa_2$, and

$$\sum_{i=1}^{n} \underline{\psi}(\theta_i; \hat{\delta}, \hat{\mu}_1, \hat{\kappa}_1, \hat{\kappa}_2) = 0, \tag{9}$$

meaning that this TMME is a Fisher consistent M-estimator.

Consider now the whole GvM_k class and the Cartesian re-parameterization

$$\begin{aligned} &\lambda_1 = \kappa_1 \cos\mu_1, \quad \lambda_2 = \kappa_1 \sin\mu_1, \\ &\lambda_3 = \kappa_2 \cos 2\mu_2 \quad \text{and} \quad \lambda_4 = \kappa_2 \sin 2\mu_2, \quad \ldots, \\ &\lambda_{2k-1} = \kappa_k \cos k\mu_k, \quad \lambda_{2k} = \kappa_k \sin k\mu_k. \end{aligned} \tag{10}$$

By defining $\underline{\lambda} = (\lambda_1, \ldots, \lambda_{2k})^{\mathrm{T}} \in \mathrm{R}^{2k}$ and

$$\underline{T}(\omega) = (\cos\omega, \sin\omega, \cos 2\omega, \sin 2\omega, \ldots, \cos k\omega, \sin k\omega)^{\mathrm{T}}, \tag{11}$$

the GvM_k density takes the $2k$-parameters canonical exponential form

$$f^*(\omega \mid \underline{\lambda}) = \exp\{\underline{\lambda}^{\mathrm{T}} \underline{T}(\omega) - K(\underline{\lambda})\}. \tag{12}$$

Thus the logarithmic likelihood is $l^*(\underline{\lambda} \mid \theta_1, \ldots, \theta_n) = \sum_{i=1}^{n} \log f^*(\theta_i \mid \underline{\lambda}) = \underline{\lambda}^{\mathrm{T}} \sum_{i=1}^{n} \underline{T}(\theta_i) - nK(\underline{\lambda})$, and the MLE of $\underline{\lambda}$, denoted $\hat{\underline{\lambda}}$, is solution in $\underline{\lambda}$ of $\mathrm{E}[\underline{T}(\theta_1)] = n^{-1} \sum_{i=1}^{n} \underline{T}(\theta_i)$, which is equivalent to solving (9) when $k = 2$. From the transformation invariance of the MLE, the MLE for $k = 2$ under the original polar parameterization μ_1, μ_2, κ_1 and κ_2 must be the polar transformations of the MLE $\hat{\underline{\lambda}}$, from where follows that the MLE is exactly the TMME above. One can also see this equivalence between MLE and TMME by comparing the first order derivatives of the logarithmic likelihood under the polar parameterization, i.e. the efficient score function (13) below, with (8). (The identity $\kappa_1 H_1(\delta, \kappa_1, \kappa_2) = -2\kappa_2[\sin 2\delta\, G_2(\delta, \kappa_1, \kappa_2) + \cos 2\delta\, H_2(\delta, \kappa_1, \kappa_2)]$ turns out to be helpful for this comparison.)

Under standard regularity conditions, the MLE has many important properties, which mainly are: strong consistency, Fisher consistency, asymptotic sufficiency, asymptotic efficiency and asymptotic normality. The efficient score function is defined as $\underline{s}(\omega; \underline{t}) = (\partial/\partial \underline{t}) \log f(\omega | \mu_1, \mu_1 - \delta, \kappa_1, \kappa_2)$, where $\underline{t} = (\delta, \mu_1, \kappa_1, \kappa_2)^{\mathrm{T}}$, and it is given by

$$\underline{s}(\omega; \delta, \mu_1, \kappa_1, \kappa_2) = \begin{pmatrix} 2\kappa_2\{\sin 2\delta A_2(\delta, \kappa_1, \kappa_2) + \cos 2\delta B_2(\delta, \kappa_1, \kappa_2) \\ \qquad - \sin 2(\omega - \mu_1 + \delta)\} \\ \kappa_1 \sin(\omega - \mu_1) + 2\kappa_2 \sin 2(\omega - \mu_1 + \delta) \\ \qquad - A_1(\delta, \kappa_1, \kappa_2) + \cos(\omega - \mu_1) \\ -\cos 2\delta A_2(\delta, \kappa_1, \kappa_2) + \sin 2\delta B_2(\delta, \kappa_1, \kappa_2) \\ \qquad + \cos 2(\omega - \mu_1 + \delta) \end{pmatrix}. \tag{13}$$

Then we have

$$\sqrt{n}(\underline{T}_n - \underline{t}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I^{-1}(\underline{t})), \quad \text{as } n \to \infty, \tag{14}$$

where $\underline{T}_n = (\hat{\delta}, \hat{\mu}_1, \hat{\kappa}_1, \hat{\kappa}_2)^{\mathrm{T}}$ is the MLE and $I(\underline{t}) = \mathrm{E}[\underline{s}(\theta_1; \underline{t}) \underline{s}^{\mathrm{T}}(\theta_1; \underline{t})]$ is the Fisher information matrix in which the expectation is taken under $\underline{t}$. Because the computation of the elements of $I(\underline{t})$ requires too many algebraic manipulations, we can rely on the empirical version given by

$$\hat{I}(\underline{T}_n) = \frac{1}{n} \sum_{i=1}^{n} \underline{s}(\theta_i; \underline{T}_n) \underline{s}^{\mathrm{T}}(\theta_i; \underline{T}_n), \tag{15}$$

in large sample inference at least. In Sect. 4 we use (15) to compute some asymptotic variance-covariance matrices.

The following formulas are important for numerical evaluations. For $\delta \in [0, \pi)$, $\kappa_1, \kappa_2 \geq 0$ and

$$S_r = \begin{cases} 1, & \text{if } \frac{r}{2} \in \mathrm{N} \backslash \{0\}, \\ 0, & \text{otherwise}, \end{cases}$$

the following expansions hold for $r = 0, 1, \ldots$,

$$\begin{aligned} G_r(\delta, \kappa_1, \kappa_2) &= I_0(\kappa_1) I_{\frac{r}{2}}(\kappa_2) \cos r\delta\, S_r + I_0(\kappa_2) I_r(\kappa_1) \\ &\quad + \sum_{j=1}^{\infty} \cos 2j\delta\, I_j(\kappa_2)\{I_{2j+r}(\kappa_1) + I_{|2j-r|}(\kappa_1)\}, \end{aligned} \tag{16}$$

and

$$\begin{aligned} H_r(\delta, \kappa_1, \kappa_2) &= -I_0(\kappa_1) I_{\frac{r}{2}}(\kappa_2) \sin r\delta\, S_r \\ &\quad + \sum_{j=1}^{\infty} \sin 2j\delta\, I_j(\kappa_2)\{I_{2j+r}(\kappa_1) \\ &\quad - I_{|2j-r|}(\kappa_1)\}. \end{aligned} \tag{17}$$

The proofs of these two useful expansions can be found in Gatto (2008, Sect. 3). From the two above expansions we can deduce that the functions G_r and H_r inherit the asymptotic behavior as $r \to \infty$ of the Bessel function I_r. From Abramowitz and Stegun (1972, 9.6.10, p. 375) it follows that $I_r(z) = (z/2)^r[r\Gamma(r)]^{-1}\{1+\mathrm{O}(r^{-1})\}$, as $r \to \infty$, which together with the Stirling expansion yields $I_r(z) = (2\pi r)^{-1/2}\{ez/(2r)\}^r\{1+\mathrm{O}(r^{-1})\}$, as $r \to \infty$. Hence I_r decreases rapidly to zero with the order r and the same holds for G_r and H_r.

Maximum likelihood estimators are used in testing problems for computing likelihood ratio test statistics. Suppose that we wish to test the null hypothesis that the independent observations $\theta_1, \ldots, \theta_n$ arise from a vM distribution, against the alternative hypothesis that this sample arises from a GvM_2 distribution. This is the testing problem of $\mathrm{H}_0 : \kappa_2 = 0$ against $\mathrm{H}_1 : \kappa_2 > 0$. Both hypotheses can be equivalently re-expressed according to the Cartesian re-reparameterization (10) with $k = 2$ as $\mathrm{H}_0 : \lambda_3 = \lambda_4 = 0$ against $\mathrm{H}_1 : \lambda_3 \neq 0$ or $\lambda_4 \neq 0$. Define $\Lambda_0 = \mathrm{R} \times \mathrm{R} \times \{0\} \times \{0\}$. Then the scaled likelihood ratio test statistic for this problem is

$$\begin{aligned} Q_n &= 2\Big\{l^*(\hat{\underline{\lambda}} \mid \theta_1, \ldots, \theta_n) - \sup_{\lambda \in \Lambda_0} l^*(\underline{\lambda} \mid \theta_1, \ldots, \theta_n)\Big\} \\ &= 2\big\{l(\hat{\mu}_1, \hat{\mu}_2, \hat{\kappa}_1, \hat{\kappa}_2 \mid \theta_1, \ldots, \theta_n) \\ &\qquad - l(\bar{\theta}_1, 0, A^{(-1)}(n^{-1}R_{1n}), 0 \mid \theta_1, \ldots, \theta_n)\big\} \\ &\xrightarrow{\mathcal{D}} \chi^2_2, \quad \text{as } n \to \infty, \end{aligned}$$

where $C_{1n} = \sum_{i=1}^n \cos\theta_i$, $S_{1n} = \sum_{i=1}^n \sin\theta_i$, $R_{1n} = (C_{1n}^2 + S_{1n}^2)^{1/2}$, $\bar{\theta}_1 = \arg\{C_{1n}, S_{1n}\}$, $l(\mu_1, \mu_2, \kappa_1, \kappa_2 \mid \theta_1, \ldots, \theta_n) = \sum_{i=1}^n \log f(\theta_i \mid \mu_1, \mu_2, \kappa_1, \kappa_2)$ and where $A^{(-1)}$ is the inverse of the function $A(\kappa) = I_1(\kappa)/I_0(\kappa)$, $\kappa \geq 0$. The accurate approximation to $A^{(-1)}$ given by Best and Fisher (1981), see also Fisher (1993, p. 51), can be used here.

2.2 Entropy and model selection

Given f and g two circular densities, we define $H(f|g) = -\int_0^{2\pi} \log g(\omega) f(\omega) d\omega$, where $0 \log 0 = 0$ and $\mathrm{dom}(f) \subset \mathrm{dom}(g)$ are assumed. We denote $f_{\underline{\eta}} = f(\cdot|\mu_1, \mu_2, \kappa_1, \kappa_2)$ the $\mathrm{GvM}_2(\mu_1, \mu_2, \kappa_1, \kappa_2)$ density, where $\underline{\eta} = (\mu_1, \kappa_1, \mu_2, \kappa_2)^{\mathrm{T}}$. Then, the differential entropy of Shannon (1948) of $f_{\underline{\eta}}$ is

$$H(f_{\underline{\eta}}|f_{\underline{\eta}}) = -\int_0^{2\pi} \log f_{\underline{\eta}}(\omega) f_{\underline{\eta}}(\omega) d\omega = \log \frac{2\pi G_0(\delta, \kappa_1, \kappa_2)}{\exp\{\kappa_1 A_1(\delta, \kappa_1, \kappa_2) + \kappa_2[\cos 2\delta A_2(\delta, \kappa_1, \kappa_2) - \sin 2\delta B_2(\delta, \kappa_1, \kappa_2)]\}}, \tag{18}$$

where $\delta = (\mu_1 - \mu_2)\mathrm{mod}\pi$ and the functions A_1, A_2 and B_2 are defined in (5) and (6), and can be evaluated with the expansions (16) and (17). $H(\delta, \kappa_1, \kappa_2) \stackrel{\mathrm{def}}{=} H(f_\eta|f_\eta)$ is location invariant in the sense that it depends on μ_1 and μ_2 through $\delta = (\mu_1 - \mu_2)\mathrm{mod}(2\pi)$ only. The entropy of the GvM_k distribution for a general k is given in Gatto (2008, Sect. 2). Note that when $\kappa_2 = 0$, the above entropy reduces to the entropy of the vM distribution $H(\cdot, \kappa_1, 0) = \log(2\pi I_0(\kappa_1)\exp\{-\kappa_1 I_1(\kappa_1)/I_0(\kappa_1)\})$. Moreover, when $\delta = \kappa_1 = 0$, $H(0, 0, \kappa_2) = H(\cdot, \kappa_2, 0)$, which confirms that the entropy of the vM distribution is equal to the entropy of the bimodal vM (vM_2) distribution

$$f(\omega \mid \cdot, \mu_2, 0, \kappa_2) = \frac{1}{2\pi I_0(\kappa_2)} \exp\{\kappa_2 \cos 2(\omega - \mu_2)\}, \tag{19}$$

for $\mu_2 \in [0, \pi)$ and $\kappa_2 \geq 0$. $H(\delta, \kappa_1, \kappa_2)$ is the maximal entropy value among all circular distributions having fixed trigonometric moments φ_1 and φ_2, where the relation between φ_1, φ_2 and $\delta, \kappa_1, \kappa_2$ is given by (7) with $r = 1, 2$, see e.g. Gatto (2008, Corollary 2.1). The entropy provides a general criterion for selecting a distribution given some partial knowledge (often based on the observed sample), according to which we should always choose distributions having maximal entropy subject to existing constraints. This is often referred to as maximum entropy principle, and the so-selected distribution is the least unsuitable one given the partial knowledge. From this and from the equivalence between TMME and MLE follows that the largest possible entropy value having trigonometric moments $\hat{\varphi}_1$ and $\hat{\varphi}_2$ is exactly $H(\hat{\delta}, \hat{\kappa}_1, \hat{\kappa}_2)$, where $\hat{\delta}$, $\hat{\kappa}_1$ and $\hat{\kappa}_2$ are the TMME or MLE of the $\mathrm{GvM}_2(\mu_1, \mu_2, \kappa_1, \kappa_2)$ distribution. Two important results in this context state that both

$$\hat{H}_n = H(\hat{\delta}, \hat{\kappa}_1, \hat{\kappa}_2) + \frac{1}{2}\frac{4}{n}$$

and

$$\tilde{H}_n = -\frac{1}{n}\sum_{i=1}^n \log f(\theta_i|\hat{\mu}_1, \hat{\mu}_1 + \hat{\delta}, \hat{\kappa}_1, \hat{\kappa}_2) + \frac{1}{2}\frac{4}{n}$$

are bias-corrected estimators of $H(\delta, \kappa_1, \kappa_2)$, in the sense that $\mathrm{E}[\hat{H}_n] = H(\delta, \kappa_1, \kappa_2) + \mathrm{o}(n^{-1})$ and $\mathrm{E}[\tilde{H}_n] = H(\delta, \kappa_1, \kappa_2) + \mathrm{o}(n^{-1})$, as $n \to \infty$, where both expectations above are taken with respect to the unknown parameters, see e.g. Zong (2006, Theorem 5.6) and Sakamoto et al. (1986, Equation 4.34) respectively. The common idea of the proofs of both results above is to consider Taylor expansions of order two, for the entropy and for the logarithmic likelihood.

In these expansions, the terms of order one vanish and the terms of order two lead to the chi-squared random variable with p degrees of freedom, where p is the number of unknown parameters in the models; $p = 4$ in our case. The bias-correction terms are directly related with the expectation of the chi-squared random variable, which is p.

The Kullback-Leibler (1951) differential information is given by

$$I(f|g) = \int_0^{2\pi} \log \frac{f(\omega)}{g(\omega)} f(\omega) d\omega = H(f|g) - H(f|f), \tag{20}$$

under the previous assumptions. $I(f|g)$ is the mean logarithmic likelihood ratio or mean information per observation of f for discriminating in favor of f against g. The Gibbs inequality tells us that $I(f|g)$ is positive semi-definite, i.e. $I(f|g) \geq 0$ for all assumed densities f and g, with equality iff $f = g$ a.e. $I(f|g)$ is sometimes called relative entropy or Kullback-Leibler distance, even though it is not a metric: it violates the symmetry and the triangle rules. From this Akaike (1973) derived the AIC for model selection, which is as follows in our setting. From (20), we see that minimizing $I(f|g)$ with respect to g corresponds to minimizing $H(f|g)$. Suppose that $f^\dagger$ is the circular density of the true model and that $g_{\underline{\nu}}$ is the density of a candidate model, having p unconstrained parameters $\underline{\nu} = (\nu_1, \ldots, \nu_p)^{\mathrm{T}}$. For some $k \geq 0$, we suppose that $g_{\underline{\nu}}$ is the GvM_k density and that $f^\dagger$ is either the GvM_k density or any other density obtained by either restricting or generalizing the GvM_k density. The restricted density $f^\dagger$ is obtained by setting some parameters of the density $g_{\underline{\nu}}$ equal to zero and the generalized density $f^\dagger$ is so that $g_{\underline{\nu}}$ results after setting some parameters of $f^\dagger$ equal to zero. For example, $f^\dagger$ is the GvM_k density, for some $k \geq 2$, and $g_{\underline{\nu}} = f_{\underline{\eta}}$ is the GvM_2 density. The density $g_{\underline{\nu}}$ provides a good approximation to $f^\dagger$ if $H(f^\dagger|g_{\underline{\nu}})$ is small. The goodness-of-fit of the maximum likelihood model can be evaluated by the expected logarithmic likelihood $-nH(f^\dagger|g_{\underline{\hat{\nu}}})$, where $\underline{\hat{\nu}}$ is the MLE of $\underline{\nu}$ based on $\theta_1, \ldots, \theta_n$ independent and with common unknown density $f^\dagger$. As this goodness-of-fit measure is random, we can evaluate it by taking the expectation with respect to $f^\dagger$. We hence obtain the mean expected logarithmic likelihood

$$\lambda_n(p) = -n\mathrm{E}[H(f^\dagger|g_{\underline{\hat{\nu}}})] \tag{21}$$

as criterion for model selection: a candidate $g_{\underline{\nu}}$ with large $\lambda_n(p)$ should be preferred. In this context the AIC statistic is defined as

$$AIC(p) = -2\sum_{i=1}^{n} \log g_{\underline{\hat{\nu}}}(\theta_i) + 2p \tag{22}$$

and $-AIC(p)/2$ provides a bias-corrected estimator of $\lambda_n(p)$, in the sense that $\mathrm{E}[-AIC(p)/2] = \lambda_n(p) + \mathrm{o}(1)$, as $n \to \infty$, see e.g. Sakamoto et al. (1986, p. 74). When $g_{\underline{\nu}} = f_{\underline{\eta}}$, i.e. when the candidate model is in the GvM_2 class, then we have

$$\begin{aligned} AIC(4) = &-2\hat{\kappa}_1 \sum_{i=1}^{n} \cos(\theta_i - \hat{\mu}_1) \\ &- 2\hat{\kappa}_2 \sum_{i=1}^{n} \cos 2(\theta_i - \hat{\mu}_1 + \hat{\delta}) \\ &+ 2n \log\{2\pi G_0(\hat{\delta}, \hat{\kappa}_1, \hat{\kappa}_2)\} + 2 \cdot 4. \end{aligned}$$

A desirable side-effect of the bias-correction term $-p$ of $-AIC(p)/2$ is to privilege candidate models with few parameters, especially with small to moderate sample sizes. Note also that only differences of AIC are meaningful, as the entropy term in (20) has been omitted in this construction. For further justifications, refer e.g. to Sakamoto et al. (1986, Chap. 4).

As mentioned, the AIC is based on the fact that $g_{\underline{\nu}}$ is a good approximation to $f^\dagger$ when $H(f^\dagger|g_{\underline{\nu}}) = I(f^\dagger|g_{\underline{\nu}}) + H(f^\dagger|f^\dagger)$ is small. If we replace the first summand $I(f^\dagger|g_{\underline{\nu}})$ by the symmetric Kullback-Leibler differential divergence $J(f^\dagger, g_{\underline{\nu}}) = I(f^\dagger|g_{\underline{\nu}}) + I(g_{\underline{\nu}}|f^\dagger)$, then we obtain an alternative measure called the total statistical entropy, namely the sum of the uncertainty due to model misuse $J(f^\dagger, g_{\underline{\nu}})$ and the uncertainty inherent in the model $H(f^\dagger|f^\dagger)$, see Zong (2006, p. 104). The ME is defined as

$$ME(p) = H(g_{\underline{\hat{\nu}}} \mid g_{\underline{\hat{\nu}}}) + \frac{3}{2}\frac{p}{n} \tag{23}$$

and it is a bias-corrected estimator to the total statistical entropy. Hence the model which minimizes $ME(p)$ should be selected, see Zong (2006, Theorems 5.10 and 5.12). As before, we suppose that the candidate $g_{\underline{\nu}}$ is a GvM_k density, $k \geq 0$, and that the true density $f^\dagger$ is either the GvM_k density or any other density obtained by restricting or by generalizing the GvM_k density. When this candidate is the GvM_2 density, then $g_{\underline{\nu}} = f_{\underline{\eta}}$ and we have $ME(4) = H(\hat{\delta}, \hat{\kappa}_1, \hat{\kappa}_2) + 3 \cdot 4/(2n)$, which can be evaluated by (18).

3 Simulation

In Sect. 3.1 we present two types of acceptance-rejection algorithms for generating pseudo-random numbers from the GvM_2 distribution: the ratio-of-uniform method and the von Neumann acceptance-rejection method. When κ_1 or κ_2 are moderate to large, then the ratio-of-uniform method is the most efficient. Otherwise, both algorithms show similar efficiency. In Sect. 3.2 we discuss the numerical determination of the sampling domains used in both algorithms. All the generation algorithms presented here are exact.

3.1 Acceptance-rejection algorithms

The generation of pseudo-random numbers from the $\mathrm{GvM}_2(\mu_1, \mu_2, \kappa_1, \kappa_2)$ distribution can be done with the ratio-of-uniforms method, which in this case yields the following algorithms: the first is a standard one and the second is an optimized version by squeezing or pretesting.

Standard ratio-of-uniforms algorithm
Step 1 Define

$$g(\omega) = \kappa_1 \cos(\omega - \mu_1) + \kappa_2 \cos 2(\omega - \mu_2) \tag{24}$$

and determine numerically

$$a = \sup_{\omega \in [0,2\pi)} \{e^{\frac{1}{2}g(\omega)}\} \quad \text{and} \quad b = \sup_{\omega \in [0,2\pi)} \{\omega e^{\frac{1}{2}g(\omega)}\}.$$

Step 2 Generate $(U, V) \sim \mathrm{Uniform}(\mathcal{P}_{a,b})$, where $\mathcal{P}_{a,b}$ denotes the body of the polygon with vertices $(0, 0)$, $(a, 0)$, (a, b) and $(b/(2\pi), b)$. Define $W = g(V/U)/2$.
Step 3 If $U \le e^W$, then consider $\theta = V/U$ as a $\mathrm{GvM}_2(\mu_1, \mu_2, \kappa_1, \kappa_2)$ pseudo-random number and stop.
Else, reject (U, V) and go to Step 2.

The computation of a can be re-expressed in terms of the search for the roots of a fourth degree polynomial and the computation of b must also be done numerically. These numerical aspects are deferred to Sect. 3.2. There are however two main advantages in redefining $b = 2\pi a$, which is an upper bound to the supremum b as given in Step 1. First, we avoid a numerical search, see Sect. 3.2. Second, while $\mathcal{P}_{a,b}$ is generally a quadrilateral, it becomes a triangle when we set $b = 2\pi a$. The simulation over this triangle can be directly done as follows: we first generate $U \sim \mathrm{Uniform}(0, a)$ and $V \sim \mathrm{Uniform}(0, 2\pi a)$ and if $V > 2\pi U$, then we replace U by $a - U$ and V by $2\pi a - V$. It would also be possible to replace a and b in Step 1 by the trivial upper bounds $a = e^{(\kappa_1+\kappa_2)/2}$ and $b = 2\pi e^{(\kappa_1+\kappa_2)/2}$.

A well-known way of decreasing the number of evaluations of the cosine and the exponential functions is by squeezing or pretesting. Under the restriction $\kappa_1 + \kappa_2 < 2$, a squeezed algorithm is the following.

Squeezed ratio-of-uniforms algorithm
Condition $\kappa_1 + \kappa_2 < 2$
Steps 1' and 2' Similar to Steps 1 and 2.
Step 3' If $U > \{1 - (\kappa_1 + \kappa_2)/2\}^{-1}$, then reject (U, V) and go to Step 2'.
Else if $U \le 1 - (\kappa_1 + \kappa_2)/2$, then consider $\theta = V/U$ as a $\mathrm{GvM}_2(\mu_1, \mu_2, \kappa_1, \kappa_2)$ pseudo-random numbers and stop.
Else if $U \le e^W$, then consider $\theta = V/U$ as a $\mathrm{GvM}_2(\mu_1, \mu_2, \kappa_1, \kappa_2)$ pseudo-random numbers and stop.
Else, reject (U, V) and go to Step 2'.

The ratio-of-uniforms method is a general method for generating random variables (see e.g. Ripley 1987) and it is a consequence of the following results. Suppose in general that h is an integrable function over a generally unbounded domain A and that $C_h = \{(u, v) \mid 0 < u \le \sqrt{h(v/u)}, v/u \in A\}$. Then C_h has a finite volume and if (U, V) is uniformly distributed over C_h, then V/U has a density over A which is proportional to h. These results can be easily shown and it is also possible to see that $C_h \subset [0, a] \times [b_-, b_+]$, where $a = \sqrt{\sup_{x \in A}\{h(x)\}}$,

$$b_+ = \begin{cases} \sqrt{\sup_{x \in A \cap \mathrm{R}_+}\{x^2 h(x)\}}, & \text{if } A \cap \mathrm{R}_+ \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases} \quad \text{and}$$

$$b_- = \begin{cases} -\sqrt{\sup_{x \in A \cap \mathrm{R}_-}\{x^2 h(x)\}}, & \text{if } A \cap \mathrm{R}_- \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

In our situation $h(\omega) = e^{g(\omega)}$, $\omega \in A = [0, 2\pi)$.

As mentioned before, the scope of the squeezed version is to minimize the number of evaluations of the cosine and of the exponential functions. The inequality $1 + w \le e^w \le (1 - w)^{-1}$ holds $\forall w \in \mathrm{R}$. It follows that if $u \le 1 + w$, then a fortiori $u \le e^w$ and we are in the acceptance region. If $u > (1 - w)^{-1}$, then a fortiori we are in the rejection region. If however none of the above conditions are fulfilled, then and only then we evaluate the exponential. If $\kappa_1 + \kappa_2 < 2$, then we may also skip the evaluation of the cosines by bounding them, which gives the squeezed version algorithm above. Simulation studies have not given evidence that the above squeezing reduces the computing time, when the operations are implemented in a vectorial way, which is essential with interpreted programming languages such as *Matlab*. In this situation, the two additional acceptances or rejections require some extra comparisons and re-indexing which apparently need more computing time than what is used for the vectorial evaluations of the cosines and of the exponentials.

A standard alternative to the ratio-of-uniforms is the von Neumann acceptance-rejection algorithm.

Von Neumann acceptance-rejection algorithm
Step 1" Define the function g as in (24) and determine, numerically, $m = \sup_{\omega \in [0,2\pi)} g(\omega)$.
Step 2" Generate $U \sim \mathrm{Uniform}(0, 2\pi)$ and $V \sim \mathrm{Uniform}(0, m)$.
Step 3" If $V \le g(U)$, then consider $\theta = U$ as a $\mathrm{GvM}_2(\mu_1, \mu_2, \kappa_1, \kappa_2)$ pseudo-random number and stop.
Else, reject U and V and go to Step 2".

The only particular part of the above algorithm is the numerical determination of the supremum m in Step 1". This is clearly similar to the determination of a in Step 1 of the ratio-of-uniforms method and it is solved in Sect. 3.2. More

efficient acceptance-rejection algorithms could be found by replacing the constant function m by a smaller envelope, i.e. by a function closer to g, while lying over g. However there are no simple envelopes for GvM_2 densities, i.e. there is not a density which, after multiplication by a constant, can cover a GvM_2 density and which allows for simple sampling.

All these algorithms are simple and do not require evaluating the normalizing constant G_0 by (16). In principle, the algorithms given here can be extended to other GvM_k distributions with $k > 2$, mainly by choosing $g(\omega) = \sum_{j=1}^{k} \kappa_j \cos j(\omega - \mu_j)$. Alternative types of generation algorithms from GvM_2 distributions are not easy to find for the following reasons: because of the complexity of the normalizing constant, because a formula for the inverse of the distribution is not available, as explained in the next paragraph, and also because there are no invariance properties which would allow to focus on the generation from a standardized version of the GvM_2 distribution.

Note that a formula for the distribution function can be obtained by Fourier series. By extending the trigonometric moment φ_r, the cosine moment γ_r and the sine moment σ_r to $r = 0, -1, \ldots$ and by noting that $\gamma_{-r} = \gamma_r$ and $\sigma_{-r} = -\sigma_r$, $r = 1, 2, \ldots$, we have from (7)

$$\begin{aligned} f(\omega|\mu_1, \mu_2, \kappa_1, \kappa_2) &= \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} \varphi_r \exp\{-\imath r\omega\} \\ &= \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{r=1}^{\infty} \gamma_r \cos r\omega + \sigma_r \sin r\omega \\ &= \frac{1}{2\pi} + \frac{1}{\pi} \sum_{r=1}^{\infty} (\cos r\omega, \sin r\omega) R(r\mu_1) \\ &\quad \times (A_r(\delta, \kappa_1, \kappa_2), B_r(\delta, \kappa_1, \kappa_2))^{\mathrm{T}}, \end{aligned}$$

where the equalities above are in the L_2 sense. It is well known that integrating term by term a Fourier series leads to a convergent series. This can be intuitively understood by the fact that integrating $\cos r\xi$ and $\sin r\xi$ yields $r^{-1} \sin r\xi$ and $-r^{-1} \cos r\xi$, meaning that the coefficients of the new series are significantly reduced in magnitude. Moreover, the new series obtained by term by term integration converges uniformly to the integrated original function. For a proof of this, refer e.g. to Pinkus and Zafrany (1997, p. 77). Thus, by integrating the GvM_2 density from 0 to $\xi \in [0, 2\pi)$, we have

$$\begin{aligned} &F(\xi|\mu_1, \mu_2, \kappa_1, \kappa_2) \\ &= \int_0^{\xi} f(\omega|\mu_1, \mu_2, \kappa_1, \kappa_2) d\omega \\ &= \frac{\xi}{2\pi} + \frac{1}{\pi} \sum_{r=1}^{\infty} \frac{1}{r} (\sin r\xi, 1 - \cos r\xi) R(r\mu_1) \\ &\quad \times (A_r(\delta, \kappa_1, \kappa_2), B_r(\delta, \kappa_1, \kappa_2))^{\mathrm{T}} \\ &= \frac{\xi}{2\pi} + \frac{1}{\pi} \sum_{r=1}^{\infty} \frac{1}{r} [A_r(\delta, \kappa_1, \kappa_2)\{\sin r(\xi - \mu_1) + \sin r\mu_1\} \\ &\quad - B_r(\delta, \kappa_1, \kappa_2)\{\cos r(\xi - \mu_1) - \cos r\mu_1\}], \end{aligned} \tag{25}$$

where the series on the right side converges uniformly to the integrated density. Note that the series in (25) is no longer a Fourier series, because $\xi/(2\pi)$ is not a term of a Fourier series. The summands of the series on the right side of (25) decrease rapidly, as they are r times smaller than the terms of a Fourier series and these are known to converge to zero (from the Riemann-Lebesgue lemma). As mentioned at the end of Sect. 2.1, the rate of decrease to zero of both functions A_r and B_r as $r \to \infty$ is comparable to the rate of decrease of the Bessel function I_r as $r \to \infty$, which is a fast rate. Consequently, only the first few summands of the series in (25) are numerically relevant. Unfortunately, this series is not practical for random variable generation, because it is not easy to obtain the inverse distribution from it. However, this new series allows e.g. to compute the probability integral transform, which is essential in many goodness-of-fit tests.

Note finally that, from the characterization property that if $\underline{X}$ is a bivariate normal vector with expectation $\underline{\nu} = (\nu_1, \nu_2)^{\mathrm{T}}$ and covariance matrix Σ, then $\arg\{\underline{X}\} \mid \|\underline{X}\| = 1$ has a $\mathrm{GvM}_2(\mu_1, \mu_2, \kappa_1, \kappa_2)$ distribution, this algorithm allows for the generation of these conditional binormal pseudo-random numbers. For the proof of this conditional representation and for the exact correspondence between $\underline{\nu}, \Sigma$ and $\mu_1, \mu_2, \kappa_1, \kappa_2$, we refer to Gatto and Jammalamadaka (2007, Characterization 2).

3.2 Numerical determination of the sampling domain

The constant a in Step 1 can be determined by searching the roots of a fourth degree polynomial. We search the solutions in $\omega \in [0, 2\pi)$ of the equation

$$\begin{aligned} &(1 - 2\sin^2\delta) \sin\omega \cos\omega - 2 \sin\delta \cos\delta \sin^2\omega \\ &\quad + \rho \sin\omega + \sin\delta \cos\delta = 0, \end{aligned} \tag{26}$$

where $\rho = \kappa_1/(4\kappa_2)$ and $\delta = (\mu_1 - \mu_2) \mathrm{mod} \pi$. In terms of $x = \sin\omega$, these extrema can be obtained by the solutions in $x \in [-1, 1]$ of the equations

$$\begin{aligned} &\pm(1 - 2\sin^2\delta) x \sqrt{1 - x^2} - 2 \sin\delta \cos\delta x^2 \\ &\quad + \rho x + \sin\delta \cos\delta = 0. \end{aligned} \tag{27}$$

Alternatively, these extrema can be found by computing the roots in $x \in [-1, 1]$ of the fourth degree polynomial

$$\begin{aligned} &x^4 - 4\rho \sin\delta \cos\delta x^3 + (\rho^2 - 1) x^2 \\ &\quad + 2\rho \sin\delta \cos\delta x + (\sin\delta \cos\delta)^2 = 0. \end{aligned} \tag{28}$$

This reformulation of (27) in terms of the fourth degree polynomial in (28) is justified as follows. We want to solve for $x \in [-1,1]$ the equations $\pm r(x) + p(x) = 0$, where $r(x) = x\sqrt{1-x^2}$ and $p(x) = c_2x^2 + c_1x + c_0$. From $\{x \in [-1,1] | p(x) + r(x) = 0 \text{ or } p(x) - r(x) = 0\} = \{x \in [-1,1] | [p(x) + r(x)][p(x) - r(x)] = 0\}$, we solve $[p(x) + r(x)][p(x) - r(x)] = 0 \Leftrightarrow$

$$x^4 + \frac{2c_2c_1}{1+c_2^2}x^3 + \frac{-1+c_1^2+2c_2c_0}{1+c_2^2}x^2 + \frac{2c_1c_0}{1+c_2^2}x + \frac{c_0^2}{1+c_2^2} = 0.$$

By inserting the values of c_0, c_1 and c_2 implied by (27) in the above polynomial, we obtain (28). The search for the roots of (28) can be easily done with e.g. *Matlab*'s routine `roots`, which re-expresses the problem into the search of the eigenvalues of the companion matrix, or with the method of Weierstrass summarized below. Then we transform these roots back to $\omega = \arcsin x$, $\pi - \arcsin x$ and retain only the values ω which satisfy (26). (As usually, $\arcsin : [-1,1] \to [-\pi/2, \pi/2]$.) We finally add modulo 2π the value of μ_1 to the values retained and evaluate $e^{g(\omega)/2}$ at these values. The largest of these evaluations yields the value of a required by Step 1.

The roots of the polynomial (28) can also be found with the method of Weierstrass, also called method of Durand-Kerner, which finds the roots of a polynomial of any degree. Consider for example the fourth degree polynomial $p(x) = x^4 + c_3x^3 + c_2x^2 + c_1x + c_0$, $x \in \mathbb{C}$. If $x_1, x_2, x_3, x_4 \in \mathbb{C}$ are the roots of p, then $p(x) = (x-x_1)(x-x_2)(x-x_3)(x-x_4)$ and it follows that

$$x_1 = x - \frac{p(x)}{(x-x_2)(x-x_3)(x-x_4)}, \tag{29}$$

for all $x \neq x_2, x_3, x_4$, i.e. x_1 is determined after one iteration. If one replaces the zeros x_2, x_3, x_4 by the approximations $x_2', x_3', x_4' \neq x_1$, then x_1 remains a fixed point of the perturbed fixed point iteration, as $x_1 = x_1 - p(x_1)/[(x_1 - x_2')(x_1 - x_3')(x_1 - x_4')]$ still holds. We have a contracting mapping around x_1 and the resulting algorithm consists of iterating (29) for all four roots until their successive differences become sufficiently small yielding, for $n = 1, 2, \ldots,$

$$x_1^{(n)} = x_1^{(n-1)} - \frac{p(x_1^{(n-1)})}{(x_1^{(n-1)} - x_2^{(n-1)})(x_1^{(n-1)} - x_3^{(n-1)})(x_1^{(n-1)} - x_4^{(n-1)})},$$

$$x_2^{(n)} = x_2^{(n-1)} - \frac{p(x_2^{(n-1)})}{(x_2^{(n-1)} - x_1^{(n)})(x_2^{(n-1)} - x_3^{(n-1)})(x_2^{(n-1)} - x_4^{(n-1)})},$$

$$x_3^{(n)} = x_3^{(n-1)} - \frac{p(x_3^{(n-1)})}{(x_3^{(n-1)} - x_1^{(n)})(x_3^{(n-1)} - x_2^{(n)})(x_3^{(n-1)} - x_4^{(n-1)})}$$

and

$$x_4^{(n)} = x_4^{(n-1)} - \frac{p(x_4^{(n-1)})}{(x_4^{(n-1)} - x_1^{(n)})(x_4^{(n-1)} - x_2^{(n)})(x_4^{(n-1)} - x_3^{(n)})}.$$

Any four complex numbers $x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, x_4^{(0)}$ are good starting points.

The determination of the constant b is more difficult as it can unfortunately not be brought into the search of the roots of a polynomial. We can apply Newton-Raphson algorithm or *Matlab*'s function `fzero` to

$$-\frac{1}{2\kappa_2} + (\omega + \mu_1)\{(1 - 2\sin^2\delta)\sin\omega\cos\omega - 2\sin\delta\cos\delta\sin^2\omega + \rho\sin\omega + \sin\delta\cos\delta\} = 0$$

with several dispersed starting points over $[0, 2\pi)$. The extrema are obtained by adding modulo 2π the value of μ_1 to the solutions obtained. We then compare the evaluations of $\omega e^{g(\omega)/2}$ at these values and we select b as the largest on these evaluations. Alternatively, one can directly maximize $\omega e^{g(\omega)/2}$ by *Matlab*'s function `fminsearch` with several starting points. A non-optimal but simple solution to this problem is to consider $b = 2\pi a$, once the value a numerically obtained.

4 Numerical study

This section provides some numerical illustrations of various results presented in Sects. 2 and 3. We consider recent real data from meteorology for which we compute the asymptotic variance-covariance matrix of the MLE under the GvM_2 model as well as the AIC and the ME for the GvM_2, vM and vM_2 models. We finally show some three-dimensional graphs of the entropy of various GvM_2 distributions.

This illustration is based on data collected by ArcticRIMS (A Regional, Integrated Hydrological Monitoring System for the Pan Arctic Land Mass, http://rims.unh.edu) and completes the study initiated by Gatto and Jammalamadaka (2007). Wind directions were measured daily from January to December 2005 on four different Pan Arctic sites at continental level: the Pan Arctic, the Europe, the Greenland and the North America basins. These four locations lead to four data sets of $n = 365$ observations. The GvM_2 distribution is fitted to these four data sets and the MLE

Table 1 AIC and ME for the GvM_2, vM and vM_2 models

	Pan Arctic	Europe	Greenland	North America
AIC GvM_2	*765.1776*	*1007.6086*	1139.5667	*944.5905*
AIC vM	1099.3159	1311.4296	1338.5079	1313.3206
AIC vM_2	905.0474	1024.8630	*1135.5678*	960.5595
ME GvM_2	*−0.2066*	*0.4027*	*0.9406*	*0.0481*
ME vM	1.5086	1.7992	1.8363	1.8067
ME vM_2	1.2425	1.4066	1.5583	1.3222

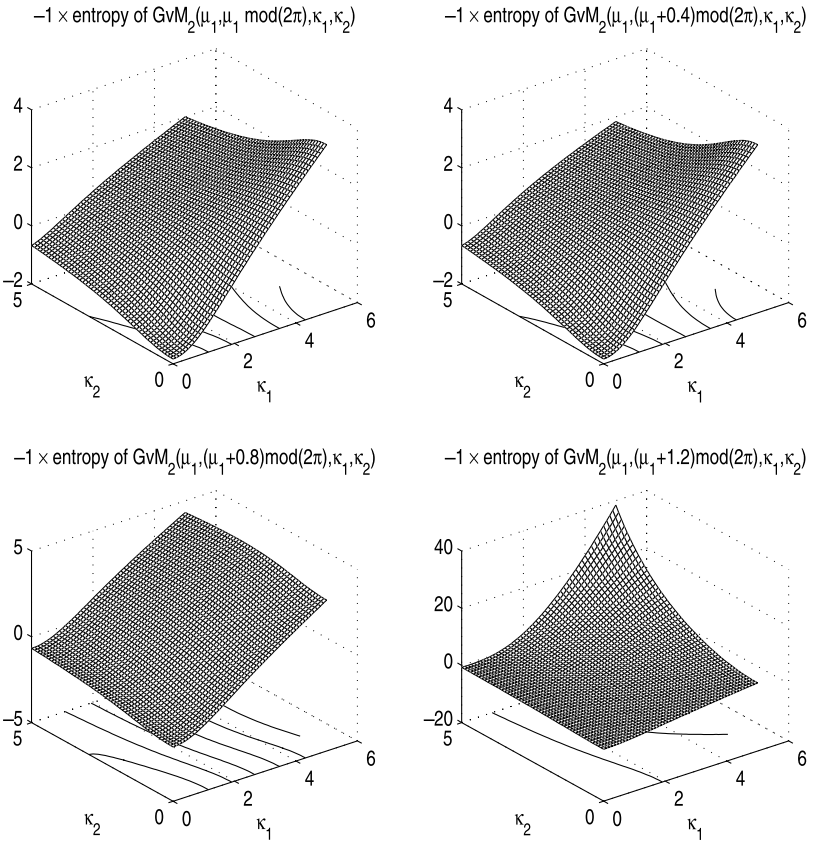

Fig. 1 Negative entropies of various GvM_2 distributions

$\underline{T}_n = (\hat{\delta}, \hat{\mu}_1, \hat{\kappa}_1, \hat{\kappa}_2)^{\mathrm{T}}$ is computed with the help of *Matlab*'s routine `fminsearch`.

The results for the four data sets are the following. For the Pan Arctic basins we have

$$\underline{T}_n = \begin{pmatrix} 0.3818 \\ 4.5055 \\ 0.8110 \\ 1.9897 \end{pmatrix} \quad \text{and}$$

$$\frac{1}{n}\hat{I}^{-1}(\underline{T}_n) = \begin{pmatrix} 0.0314 & 0.0296 & 0.0073 & 0.0031 \\ 0.0296 & 0.0285 & 0.0067 & 0.0030 \\ 0.0073 & 0.0067 & 0.0075 & -0.0001 \\ 0.0031 & 0.0030 & -0.0001 & 0.0047 \end{pmatrix}.$$

For the Europe basins we have

$$\underline{T}_n = \begin{pmatrix} 0.2384 \\ 4.2330 \\ 0.2781 \\ 1.6028 \end{pmatrix} \quad \text{and}$$

$$\frac{1}{n}\hat{I}^{-1}(\underline{T}_n) = \begin{pmatrix} 0.1824 & 0.1811 & 0.0094 & -0.0002 \\ 0.1811 & 0.1807 & 0.0092 & -0.0001 \\ 0.0094 & 0.0092 & 0.0043 & -0.0004 \\ -0.0002 & -0.0001 & -0.0004 & 0.0042 \end{pmatrix}.$$

For the Greenland basins we have

$$\underline{T}_n = \begin{pmatrix} 0.6936 \\ 4.7875 \\ 3.7756 \cdot 10^{-5} \\ 1.2119 \end{pmatrix} \quad \text{and}$$

$$\frac{1}{n}\hat{I}(\underline{T}_n) = \begin{pmatrix} 2.0647 & -2.0646 & -0.0923 & 0.0240 \\ -2.0646 & 2.0646 & 0.0923 & -0.0240 \\ -0.0923 & 0.0923 & 0.5471 & -0.0591 \\ 0.0240 & -0.0240 & -0.0591 & 0.6486 \end{pmatrix}. \quad (30)$$

The condition number (the ratio of the largest singular value to the smallest) of this matrix is $1.9442 \cdot 10^{10}$, meaning that this matrix is too close to singularity, hence not accurately invertible and we cannot obtain the asymptotic variance-covariance matrix. This is a consequence of indeterminacy brought by the very small value of $\hat{\kappa}_1$: as already mentioned near to end of Sect. 1, when $\kappa_1 = 0$ then μ_1 is indeterminate and only one single location parameter remains relevant. The two first rows or columns of the empirical Fisher information matrix in (30) are indeed practically equal, up to the sign. For the North America basins we have

$$\underline{T}_n = \begin{pmatrix} 0.8212 \\ 4.9710 \\ 0.3440 \\ 1.8601 \end{pmatrix} \quad \text{and}$$

$$\frac{1}{n}\hat{I}^{-1}(\underline{T}_n) = \begin{pmatrix} 0.0850 & 0.0836 & 0.0206 & 0.0025 \\ 0.0836 & 0.0830 & 0.0201 & 0.0024 \\ 0.0206 & 0.0201 & 0.0111 & 0.0005 \\ 0.0025 & 0.0024 & 0.0005 & 0.0038 \end{pmatrix}.$$

In Table 1 we give the AIC and the ME for the GvM_2 model, see (22) and (23), and also for the vM and vM_2 submodels. With respect to the AIC, the GvM_2 model is alway the best of the three, excepting for Greenland basin, where the vM_2 is slightly better. This result is in accordance with the previous remark that $\hat{\kappa}_1$ is very close to zero and with the histogram of the data in Gatto and Jammalamadaka (2007, Fig. 3), which shows two identical modes. The GvM_2 model is always the best with respect to the ME, even though the smallest difference in ME between the GvM_2 and vM_2 is for the Greenland basin. The best values of AIC and ME in Table 1 are given in italic.

We conclude this section with four graphical illustrations of the entropy of the GvM_2 distribution, which demonstrate the effectiveness of the entropy formula (18) based on the Fourier expansions (16) and (17). Figure 1 provides the three-dimensional graphs of $-H(\delta, \kappa_1, \kappa_2)$ for $\delta = 0$, 0.4, 0.8, 1.2 and for values of κ_1 and κ_2 in the interval $[0, 5]$, μ_1 being irrelevant for the entropy. We can recognize the well-known result that the circular uniform distribution ($\kappa_1 = \kappa_2 = 0$) maximizes the entropy.

The four samples and the *Matlab*'s programs used for this article are available at http://www.stat.unibe.ch/~gatto.

Acknowledgements The author thanks the Editor, the Associate Editor and two anonymous referees for thoughtful comments and corrections which improved the quality of this article.

References

Abramowitz, M., Stegun, I.E.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th edn. Dover, New York (1972), originally published by the National Bureau of Standards, USA, 10th edn.

Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (eds.) Second International Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest (1973)

Best, D.J., Fisher, N.I.: The bias of the maximum likelihood estimators of the von Mises-Fisher concentration parameters. Commun. Stat. B, Simul. Comput. **10**, 493–502 (1981)

Cheng, R.C.H., Amin, N.A.K.: Estimating parameters in continuous univariate distributions with shifted origin. J. R. Stat. Soc. B **54**, 394–403 (1983)

Cheng, R.C.H., Iles, T.C.: Corrected maximum likelihood in non-regular problems. J. R. Stat. Soc. B **49**, 95–101 (1987)

Cheng, R.C.H., Traylor, L.: Non-regular maximum likelihood problems, with discussion. J.R. Stat. Soc. B **57**, 3–44 (1995)

Cox, D.R., Hinkley, D.V.: Theoretical Statistics. Chapman and Hall, Cambridge (1974)

Fisher, N.I.: Statistical Analysis of Circular Data. Cambridge University Press, Cambridge (1993)

Gatto, R.: Information theoretic results for circular distributions. Technical Report, Institute of Mathematical Statistics and Actuarial Science, University of Bern (2008) under revision for Statistics

Gatto, R., Jammalamadaka, S.R.: Inference for wrapped symmetric α-stable circular models. Sankhyā A **65**, 333–355 (2003)

Gatto, R., Jammalamadaka, S.R.: The generalized von Mises distribution. Stat. Method. **4**, 341–353 (2007)

Hartigan, J.A.: A failure of likelihood asymptotics for the mixture model. In: Le Cam, L., Olshen, R.A. (eds.) Proceedings of the Berkeley Symposium in Honor of J. Neyman and J. Kiefer, vol. 2, pp. 807–810. Wadsworth, New York (1985)

Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)

Maksimov, V.M.: Necessary and sufficient statistics for the family of shifts of probability distributions on continuous bicompact groups. Teor. Veroyat. Primen. **12**, 307–321 (1967) (in Russian); Theory Probab. Appl. **12**, 267–280 (English translation)

Pinkus, A., Zafrany, S.: Fourier Series and Integral Transforms. Cambridge University Press, Cambridge (1997)

Ranneby, B.: The maximum spacing method: an estimation method related to the maximum likelihood method. Scand. J. Stat. **11**, 93–112 (1984)

Ripley, B.D.: Stochastic Simulation. Wiley, New York (1987)

Sakamoto, Y., Ishiguro, M., Kitagawa, G.: Akaike Information Criterion Statistics. Reidel, Norwell (1986)

Shannon, C.E.: The mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)

Spurr, B.D., Koutbeiy, M.A.: A comparison of various methods for estimating the parameters in mixtures of von Mises distributions. Commun. Stat. Simul. Comput. **20**, 725–741 (1991)

Titterington, D.M., Smith, A.F.M., Makov, U.E.: Statistical Analysis of Finite Mixture Distributions. Wiley, New York (1985)

Zong, Z.: Information-Theoretic Methods for Estimating Complicated Probability Distributions. Elsevier, Amsterdam (2006)