Applied Mathematics and Computation 218 (2012) 11880–11890

Contents lists available at SciVerse ScienceDirect

Applied Mathematics and Computation

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

High-order parameter approximation for von Mises–Fisher distributions

Heping Song [a,b], Jun Liu [a], Guoli Wang [a,*]

[a] School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510006, China
[b] School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China

ARTICLE INFO

Keywords:
von Mises–Fisher distribution
Langevin distribution
Parameter estimation
Halley method
Modified Bessel function

ABSTRACT

This paper concerns the issue of the maximum-likelihood estimation (MLE) for the concentration parameters of the von Mises–Fisher (vMF) distributions, which are crucial to directional data analysis. In particular, we study the numerical approximation approach for solving the implicit nonlinear equation arising from building the MLE of the concentration parameter κ of vMF distributions. In addition, we address the implementation of $I_s(x)$, the modified Bessel function of the first kind, which is the most time-consuming and fundamental ingredient in the proposed approximation scheme of the MLE for κ. The main contribution of this paper is two fold. The first is to present a two-steps Halley based method for exploring a high-order approximation of the MLE for κ, which can significantly contribute to the improvement of estimation accuracy. The second is to develop a novel approach for the implementation of $I_s(x)$, which can make the substantial improvement of computation efficiency for computing the MLE approximation for κ. The numerical experiments were conducted to compare the proposed schemes with those in the existing works by Tanabe et al. [1] and Sra [2]. The experimental results show that, given the same amount of computation as their methods, the proposed high-order scheme can achieve much more accurate approximations while our implementation of $I_s(x)$ is preferable yet desirable for high dimensional applications.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

The von Mises–Fisher (vMF) distribution has attracted considerable attention in various data analysis applications, such as text mining and gene expression analysis [3], due to its potential clustering capability for directional data. The estimation of the concentration parameter κ is a specific important problem in using vMF distributions for data mining tasks. Maximum-likelihood estimation (MLE) techniques are the mostly used approaches to estimate parameters for vMF distributions. However, seeking the MLE of κ is not a trivial task, as it is inevitably associated with an implicit nonlinear equation that challenges the existing numerical techniques in pursuit of accurate yet efficient estimations. This paper aims to find a high accuracy estimation yet computational efficient approach of building the approximation of the MLE for κ, which is a followup to the works of Tanabe et al. [1] and Sra [2].

From the numerical computation perspective, the fixed-point iteration paradigm is an effective way of building the numerical approximations of the roots of nonlinear equations. Under this paradigm, the work [1] utilizes the inverse of modified Bessel function ratio to construct the iterated function used in fixed-point iteration. In doing this, the upper and lower bounds of the MLE for κ derived play an important role in performing a linear interpolation at each iteration. Sra [2] proposed an alternative approximation using two-steps of Newton method, in which the iterated functions were derived from Taylor series expansion. Sra's approximation scheme has the close computational cost as Tanabe's method.

* Corresponding author at: School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China.
E-mail addresses: songhp@ujs.edu.cn (H. Song), stp04lj@mail2.sysu.edu.cn (J. Liu), isswgl@mail.sysu.edu.cn (G. Wang).

0096-3003/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amc.2012.05.050

This paper presents a two-steps Halley based scheme in performing fixed-point iterations for achieving a high-order approximation of the MLE for κ. Compared with the first order Householder's approximation paradigm used in Sra's scheme [2], our proposed scheme explores the second order Householder's approximation paradigm [4], which can contribute to the improvement of estimation accuracy. In addition, we address the implementation of $I_s(x)$, the modified Bessel function of the first kind, which is the most time-consuming and fundamental ingredient in the proposed approximation scheme of the MLE for κ. A novel approach for the implementation of $I_s(x)$ is developed to improve computation efficiency in computing the MLE approximation for κ. The experimental results show that our scheme performs two iterations to remain competitive in terms of running times with Tanabe's method [1] and has better convergence than Sra's method [2] while exhibiting more accurate approximation. We can conclude that, given the same amount of computation as their methods, the proposed high-order approximation scheme can achieve much more accurate approximations while our implementation of $I_s(x)$ is preferable yet desirable for high dimensional applications.

The remainder of this paper is organized as follows. In Section 2, we discuss the vMF distribution and maximum likelihood estimation of its parameters. Then we outline various approximations and present our approach in Section 3. Section 4 introduces the implementation of $I_s(x)$. Numerical experiments were conducted in Section 5. Finally, we conclude this paper in Section 6.

2. Preliminaries

In this section we discuss the von Mises–Fisher distribution and maximum likelihood estimation of its parameters.

2.1. The von Mises–Fisher (vMF) distribution

The vMF distribution is one of the simplest generative models for directional data. The probability density function of the von Mises–Fisher distribution for the random d-dimensional unit vector x (i.e., $x \in \mathbb{R}^d$ and $\|x\|_2 = 1$, or equivalently $x \in \mathbb{S}^{d-1}$, and $\mathbb{S}^{d-1}$ is the d dimensional unit hypersphere.) is given by

$$p(x|\mu,\kappa) = c_d(\kappa)e^{\kappa\mu^T x}, \tag{1}$$

where $\kappa \geqslant 0, \|\mu\|_2 = 1$ and $d \geqslant 2$. The normalization constant $c_d(\kappa)$ is given by

$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}, \tag{2}$$

where $I_d(\cdot)$ denotes the modified Bessel function of the first kind and order d. For a real number d, the function can be computed using (see [5, pp. 77], [6, Eq. (9.6.10)])

$$I_d(x) = \sum_{k=0}^{\infty} \frac{1}{\Gamma(k+1)\Gamma(d+k+1)} \left(\frac{x}{2}\right)^{2k+d}, \tag{3}$$

where $\Gamma(x)$ is the well-known Gamma function.

The parameters μ and κ are called the mean direction and concentration parameter, respectively. The κ characterizes the dispersion around the mean direction, analogous to covariance for the multivariate Gaussian distribution. The greater the value of κ is, the higher the concentration of the distribution around the mean direction is. In particular, the distribution is unimodal for $\kappa > 0$. $p(x|\mu,\kappa)$ is uniform on the hypersphere $\mathbb{S}^{d-1}$ for $\kappa = 0$, and tends to a point density for $\kappa \to \infty$. If $d = 2$, the one-dimensional unit circle is called the von Mises distribution. If $d = 3$, the distribution is called the Fisher distribution.

The vMF distribution is thought as the natural parametric distributions to directional data, and is akin to the Gaussian distribution for multivariate data in $\mathbb{R}^d$ [7]. The vMF distribution was introduced by Watson and Williams [8], and discussed in detail in the monograph of Mardia and Jupp [9] and the technical report of Dhillon and Sra [10]. The vMF distribution also refers to Langevin distribution in some literature [7,11,12].

2.2. Maximum likelihood estimates

In this subsection, we derive briefly maximum likelihood estimates (MLE) for the parameters of a vMF distribution. Let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ be the set of data drawn from a vMF distribution, and it is independent, identically distributed (iid) with unknown parameters (μ,κ). As a function of (μ,κ) with $x_1, x_2, \ldots, x_n$ fixed, the log-likelihood function is given by

$$L(\mathcal{X}|\mu,\kappa) = n \ln c_d(\kappa) + \sum_i \kappa\mu^T x_i. \tag{4}$$

To obtain the MLE, we have to maximize (4), subject to the constrains $\mu^T\mu = 1$ and $\kappa \geqslant 0$. Let $\hat{\mu}$ and $\hat{\kappa}$ denote the MLE of μ and κ, respectively. We obtain the following equations

$$\hat{\mu} = \frac{\sum_i x_i}{||\sum_i x_i||_2}, \tag{5}$$

$$A_d(\hat{\kappa}) = -\frac{c_d'(\hat{\kappa})}{c_d(\hat{\kappa})} = \frac{I_{d/2}(\hat{\kappa})}{I_{d/2-1}(\hat{\kappa})} = \frac{||\sum_i x_i||_2}{n} = \overline{R}, \tag{6}$$

i.e.

$$\hat{\kappa} = A_d^{-1}(\overline{R}). \tag{7}$$

The detailed derivation of these MLE equations can be found in [3,9]. As seen form Eq. (6), since $A_d(\hat{\kappa})$ is a ration of modified Bessel functions, there is no analytical solution to estimate the κ. We have to seek for asymptotic or numerical approximations of κ by solving $f(\kappa) = 0$, where $f(\kappa)$ is defined as

$$f(\kappa) = A_d(\kappa) - \overline{R}. \tag{8}$$

Several approximations have been discussed in computational statistics community. Hill [13] presented an approximation of the MLE of κ for the two or three dimensional vMF distributions. Mardia and Jupp [9] provided the asymptotic approximations for low dimensional cases $\overline{R} \geqslant 0.9$ or $\overline{R} < 0.05$. Watamori [12] derived the MLE of κ in the forms of asymptotic expansions, and presented a survey of the earlier works. On the other hand, Schou [14] provided the marginal maximum likelihood estimator of κ for $\overline{R} \to 1$. Dowe et al. [15] applied the information theoretic Minimum Message Length (MML) principle to estimate κ for $d = 3$.

In the next section, we discuss various approximations of MLE for κ for the general cases.

3. Parameter approximations

This section introduces several parameter approximations of κ by solving the nonlinear equation $f(\kappa) = 0$. Firstly, Tanabe et al. [1] derived the theoretical upper and lower bounds of the approximation range, and offered an approximation based on a fixed-point linear interpolation method. Secondly, these bounds show the validity of an empirical approximation presented by Banerjee et al. [3], which was derived by truncated the continued fraction representation of the modified Bessel function ratio $A_d(\kappa)$ and then added a correction term. Finally, The truncated methods proposed by Sra [2] and this paper are presented. All methods presented here give different level of approximation accuracy.

3.1. Linear interpolation approximation

Tanabe et al. [1] derived the upper and lower bounds of the MLE $\hat{\kappa}$, which are given by

$$\kappa_l = \frac{\overline{R}(d-2)}{1-\overline{R}^2} \leqslant \hat{\kappa} \leqslant \kappa_u = \frac{\overline{R}d}{1-\overline{R}^2}. \tag{9}$$

From Eq. (6) , Tanabe et al. [1] proposed a fixed point iteration function to obtain the MLE for κ, which is given by

$$\Phi(\kappa) = \overline{R}\kappa A_d(\kappa)^{-1}. \tag{10}$$

Using the intersection of $\Phi(\kappa) = \overline{R}\kappa A_d(\kappa)^{-1}$ with $\Phi(\kappa) = \kappa$ and the bounds κ_l, κ_u, Tanabe et al. [1] presented the linear interpolation approximation, which is given by

$$\kappa_T = \frac{\kappa_l\Phi(\kappa_u) - \kappa_u\Phi(\kappa_l)}{(\Phi(\kappa_u) - \Phi(\kappa_l)) - (\kappa_u - \kappa_l)}. \tag{11}$$

3.2. Heuristic approximation

Banerjee et al. [3] provided an empirical approximation of $\hat{\kappa}$, which is given by

$$\kappa_B = \frac{\overline{R}(d-\overline{R}^2)}{1-\overline{R}^2}. \tag{12}$$

This approximation can be interpreted as a correction version of the upper bound κ_u.

3.3. Truncated Newton approximation

Sra [2] proposed an approximation by performing two steps of Newton method, which is given by

$$\kappa_1 = \kappa_0 - \frac{f(\kappa_0)}{f'(\kappa_0)}, \tag{13}$$

$$\kappa_N = \kappa_1 - \frac{f(\kappa_1)}{f'(\kappa_1)}, \tag{14}$$

where

$$\kappa_0 = \kappa_B, \tag{15}$$

$$f'(\kappa) = 1 - A_d(\kappa)^2 - \frac{d-1}{\kappa}A_d(\kappa). \tag{16}$$

3.4. Truncated Halley approximation

As shown in Sra's [2] comparison experiments, the significant improvement in accuracy of approximation using truncated Newton method has not achieved much order of magnitude, especially when the variance propagated by numerical errors. Motivated by Sra [2], we exploit a high-order method to obtain much more accurate approximation. We make use of the fact that

$$f''(\kappa) = 2A_d(\kappa)^3 + \frac{3(d-1)}{\kappa}A_d(\kappa)^2 + \frac{d^2-d-2\kappa^2}{\kappa^2}A_d(\kappa) - \frac{d-1}{\kappa}, \tag{17}$$

while obtaining the Halley iterations for solving $f(\kappa) = 0$. We present a new approximation by computing two Halley steps

$$\kappa_1 = \kappa_0 - \frac{2f(\kappa_0)f'(\kappa_0)}{2f'(\kappa_0)^2 - f(\kappa_0)f''(\kappa_0)}, \tag{18}$$

$$\kappa_H = \kappa_1 - \frac{2f(\kappa_1)f'(\kappa_1)}{2f'(\kappa_1)^2 - f(\kappa_1)f''(\kappa_1)}, \tag{19}$$

where $\kappa_0, f'(\kappa)$ and $f''(\kappa)$ are defined in (15)–(17), respectively.

Remarks.

1. Let $d = 2$, the lower bound κ_l is equal to zero, so the Eq. (11) would be a constant 0. The Tanabe's approximation (11) is not available.
2. The Banerjee's approximation (12) has important difference from other approximations that it does not need additional evaluation of $A_d(\kappa)$. The basic operation of computing (11), (14) and (19), which is the most time-consuming, is the evaluation of $A_d(\kappa)$. To competitive in terms of running time with (11), (14) and (19) also require only two calls of evaluation of $A_d(\kappa)$. When the computational cost of $\overline{R}$ is larger than the cost of $A_d(\kappa)$, the approximation (19) is recommended, otherwise the approximation (12) is preferable.
3. The accuracy of different approximations is an academic concern, as also discussed by Tanabe et al. [1] and Sra [2]. However, it is a natural choice to exploit an approximation as accurate as possible, especially when the computational cost does not matter.
4. The Newton method and Halley method were derived from Taylor-series expansion, and the Halley updates made use of the Newton iteration formula. In numerical analysis, Halley method has a cubic convergence [16], while the order of convergence of Newton method is second [17,18]. As shown in our experiments, the truncated Halley method (19) exhibits the most accurate approximation with the same of two evaluations of $A_d(\kappa)$ for Tanabe's method (11) and truncated Newton method (14). Another cubic convergent method is Euler method, but it can be numerically unstable [17]. Higher-order Householder's methods are practically not used due to more expensive computation. Recently, higher-order convergent methods for solving nonlinear equation were introduced in [18–20]. However, these methods require much more additional computation, and are deemed impractical for our concerned applications.

4. Implementing $I_s(x)$

As discussed in previous section, the basic operation of computing approximations (11), (14) and (19) is the evaluation of the modified Bessel function ratio $A_d(\kappa)$. In this section, we discuss the implementation of $I_s(x)$, the modified Bessel function of the first kind, which is the most time-consuming and important computation in vMF distributions related applications. Consequently, the modified Bessel function ratio can be evaluated by computing Bessel function $I_s(x)$ and dividing.

Recall that the modified Bessel function of the first kind [6, Eq. (9.6.10)] is defined as

$$I_s(x) = \left(\frac{x}{2}\right)^s \sum_{k=0}^{\infty} \frac{(x^2/4)^k}{\Gamma(s+k+1)k!}. \tag{20}$$

Using recurrence relation of Gamma function, one can show that

$$\Gamma(x+1) = x\Gamma(x). \tag{21}$$

Then we have obtained that

$$I_s(x) = \frac{(x/2)^s}{\Gamma(s+1)}\left[1 + \sum_{k=1}^{\infty} \frac{(x^2/4)^k}{(s+1)(s+2)\cdots(s+k)k!}\right]. \tag{22}$$

The Gamma function $\Gamma(s+1)$ in (22) can be computed by partial fraction-type approximation. As Shown in Pugh's thesis [21], there are three, Lanczos [22], Spouge [23] and Stirling, formula for approximating Gamma function. For the evaluation of $\Gamma(s+1)$ with a uniformly bounded error, the Lanczos method is the best candidate (see [21, Section 9.4]). We adopt the formula used in "Numerical Recipes in C" [24] with an error that is smaller than $|\varepsilon| < 2 \times 10^{-10}$, and is given by

$$\Gamma(s+1) = \sqrt{2\pi}(s+5.5)^{(s+0.5)} e^{(-s-5.5)}\left(c_0 + \sum_{n=1}^{6} \frac{c_n}{s+n} + \varepsilon\right), \tag{23}$$

where c_i is the coefficients

$$\begin{aligned} c_0 &= 1.000000000190015,\\ c_1 &= 76.18009172947146,\\ c_2 &= -86.50532032941677,\\ c_3 &= 24.01409824083091,\\ c_4 &= -1.231739572450155,\\ c_5 &= 1.208650973866179 \times 10^{-3},\\ c_6 &= -5.395239384953 \times 10^{-6}. \end{aligned} \tag{24}$$

From the series expansion formula (22), the expression in the square brackets is the confluent hypergeometric limit function [25, p. 359]. We make use of the fact [5, p. 16] that the ratio of the kth term to the $(k-1)$st term is

$$\frac{x^2/4}{(s+k)k}, \tag{25}$$

and this tends to zero as $k \to \infty$, for all values of s and x. The summation over k can be approximated by truncating the power-series in (22). For high dimensional applications, the floating point arithmetic may introduce overflow errors, and this becomes a serious concern in Gamma function calculations. To prevent the overflow errors, it is better to compute logarithmic $I_s(x)$. Thus we obtain Algorithm 1 for implementing $\ln I_s(x)$.

Algorithm 1. Computing $\ln I_s(x)$ via truncated power-series

Input:

s, x: positive real numbers
$c_0, c_1, \ldots, c_6$: the coefficients
ε: convergence tolerance

Output:

logb: the approximation to $\ln I_s(x)$

$t_1 \leftarrow c_0 + \frac{c_1}{s+1} + \frac{c_2}{s+2} + \cdots + \frac{c_6}{s+6}$
$t_2 \leftarrow (s+0.5)\ln(s+5.5) - s - 5.5$
$t_1 \leftarrow s\ln(0.5x) - 0.5\ln(2\pi) - t_2 - \ln t_1$
$R \leftarrow 1.0$
$M \leftarrow 1.0$
$k \leftarrow 1$
while 1 **do**
  $R \leftarrow R\frac{0.25x^2}{(s+k)k}$
  $M \leftarrow M + R$
  if $R/M < \varepsilon$ **then**
    break
  end if
  $k \leftarrow k + 1$
end while
return $logb \leftarrow t_1 + \ln M$

Table 1
The absolute errors $|\hat{\kappa} - \kappa_{true}|$ for different approximations of κ. A "N/A" indicates that the approximation is not available.

(d, κ_{true})	Banerjee (12)	Tanabe (11)	Newton (14)	Halley (19)
(2,10)	4.20E−01	N/A	3.43E−05	**9.27E−22**
(2,50)	4.85E−01	N/A	4.49E−07	**1.81E−33**
(2,100)	4.92E−01	N/A	5.93E−08	**2.57E−38**
(2,500)	4.98E−01	N/A	4.95E−10	**1.60E−49**
(2,1000)	4.99E−01	N/A	6.22E−11	**2.43E−54**
(2,5000)	5.00E−01	N/A	4.99E−13	**1.58E−65**
(2,10,000)	5.00E−01	N/A	6.25E−14	**2.42E−70**
(10,10)	1.63E−01	1.50E−02	2.56E−07	**6.23E−21**
(10,50)	4.10E−01	3.20E−02	2.01E−07	**1.65E−28**
(10,100)	4.54E−01	3.34E−02	4.01E−08	**5.28E−33**
(10,500)	4.91E−01	3.42E−02	4.58E−10	**6.01E−44**
(10,1000)	4.95E−01	3.43E−02	5.98E−11	**9.80E−49**
(10,5000)	4.99E−01	3.44E−02	4.96E−13	**6.78E−60**
(10,10,000)	5.00E−01	3.44E−02	6.22E−14	**1.04E−64**
(50,10)	6.45E−03	1.76E−05	1.83E−15	**1.47E−34**
(50,50)	1.70E−01	2.15E−03	2.11E−09	**9.28E−27**
(50,100)	2.96E−01	3.75E−03	4.80E−09	**6.68E−29**
(50,500)	4.52E−01	5.05E−03	3.10E−10	**2.67E−38**
(50,1000)	4.76E−01	5.19E−03	4.94E−11	**6.38E−43**
(50,5000)	4.95E−01	5.29E−03	4.77E−13	**5.92E−54**
(50,10,000)	4.98E−01	5.30E−03	6.11E−14	**9.43E−59**
(100,10)	9.26E−04	3.54E−07	1.69E−20	**3.44E−44**
(100,50)	5.97E−02	2.64E−04	6.35E−12	**4.21E−30**
(100,100)	1.70E−01	1.03E−03	2.64E−10	**3.39E−29**
(100,500)	4.08E−01	2.30E−03	1.88E−10	**2.76E−36**
(100,1000)	4.52E−01	2.44E−03	3.87E−11	**1.11E−40**
(100,5000)	4.90E−01	2.55E−03	4.55E−13	**1.51E−51**
(100,10,000)	4.95E−01	2.56E−03	5.96E−14	**2.51E−56**
(500,10)	7.92E−06	2.53E−11	6.70E−33	**3.06E−68**
(500,50)	9.54E−04	7.27E−08	1.59E−22	**1.21E−49**
(500,100)	6.84E−03	1.82E−06	2.50E−18	**2.56E−42**
(500,500)	1.71E−01	1.99E−04	2.11E−12	**8.23E−35**
(500,1000)	2.96E−01	3.49E−04	4.70E−12	**6.38E−37**
(500,5000)	4.52E−01	4.76E−04	3.08E−13	**2.97E−46**
(500,10,000)	4.75E−01	4.90E−04	4.92E−14	**7.27E−51**
(1000,10)	9.96E−07	3.98E−13	2.63E−38	**9.51E−79**
(1000,50)	1.23E−04	1.22E−09	7.54E−28	**6.03E−60**
(1000,100)	9.58E−04	3.65E−08	2.03E−23	**4.93E−52**
(1000,500)	6.06E−02	2.59E−05	6.71E−15	**4.33E−38**
(1000,1000)	1.71E−01	9.93E−05	2.64E−13	**3.19E−37**
(1000,5000)	4.07E−01	2.23E−04	1.86E−13	**2.85E−44**
(1000,10,000)	4.52E−01	2.37E−04	3.85E−14	**1.17E−48**
(5000,10)	7.99E−09	2.56E−17	7.05E−51	**3.40E−103**
(5000,50)	9.99E−07	7.99E−14	2.14E−40	**2.52E−84**
(5000,100)	7.98E−06	2.55E−12	6.97E−36	**3.31E−76**
(5000,500)	9.61E−04	7.32E−09	1.65E−25	**1.30E−57**
(5000,1000)	6.88E−03	1.83E−07	2.58E−21	**2.71E−50**
(5000,5000)	1.71E−01	1.98E−05	2.11E−15	**8.13E−43**
(5000,10,000)	2.96E−01	3.46E−05	4.69E−15	**6.35E−45**
(10,000,10)	1.00E−09	4.00E−19	2.69E−56	**9.96E−114**
(10,000,50)	1.25E−07	1.25E−15	8.22E−46	**7.41E−95**
(10,000,100)	9.99E−07	3.99E−14	2.69E−41	**9.89E−87**
(10,000,500)	1.24E−04	1.22E−10	7.69E−31	**6.26E−68**
(10,000,1000)	9.61E−04	3.66E−09	2.07E−26	**5.11E−60**
(10,000,5000)	6.07E−02	2.59E−06	6.75E−18	**4.34E−46**
(10,000,10,000)	1.71E−01	9.89E−06	2.63E−16	**3.17E−45**
(100,000,10)	1.00E−12	4.00E−25	2.70E−74	**1.00E−148**
(100,000,50)	1.25E−10	1.25E−21	8.24E−64	**7.45E−130**
(100,000,100)	1.00E−09	4.00E−20	2.70E−59	**9.99E−122**
(100,000,500)	1.25E−07	1.25E−16	8.23E−49	**7.43E−103**
(100,000,1000)	1.00E−06	4.00E−15	2.69E−44	**9.93E−95**
(100,000,5000)	1.24E−04	1.22E−11	7.70E−34	**6.29E−76**
(100,000,10,000)	9.61E−04	3.66E−10	2.07E−29	**5.13E−68**

(*continued on next page*)

Table 1 (*continued*)

(d, κ_{true})	Banerjee (12)	Tanabe (11)	Newton (14)	Halley (19)
(1,000,000,10)	1.00E−15	4.00E−31	2.70E−92	**1.00E−183**
(1,000,000,50)	1.25E−13	1.25E−27	8.24E−82	**7.45E−165**
(1,000,000,100)	1.00E−12	4.00E−26	2.70E−77	**1.00E−156**
(1,000,000,500)	1.25E−10	1.25E−22	8.24E−67	**7.45E−138**
(1,000,000,1000)	1.00E−09	4.00E−21	2.70E−62	**1.00E−129**
(1,000,000,5000)	1.25E−07	1.25E−17	8.23E−52	**7.44E−111**
(1,000,000,10,000)	1.00E−06	4.00E−16	2.69E−47	**9.93E−103**

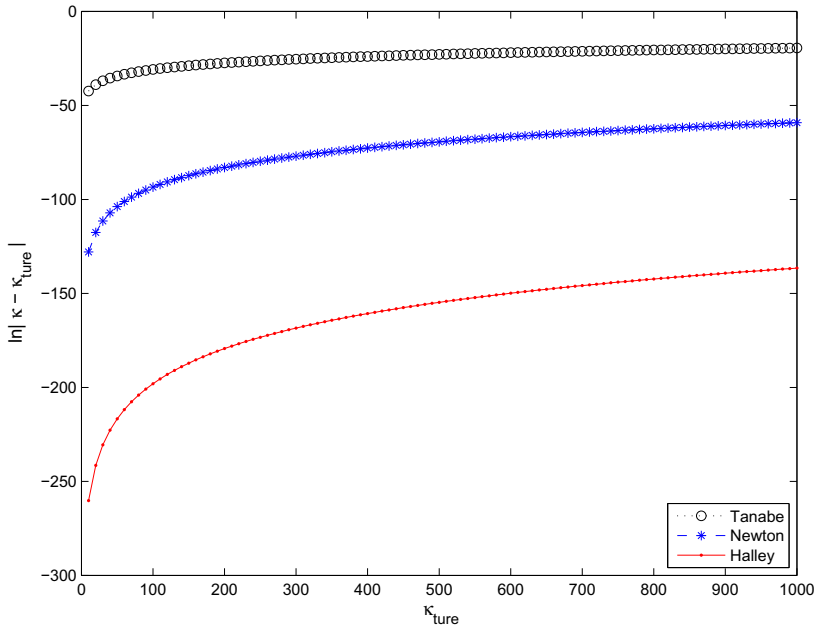

Fig. 1. Logarithmic absolute errors of approximations with varying κ from 10 to 1000 and fixed d=10,000.

Remarks

1. The most important difference between Algorithm 1 and the method proposed by Tanabe et al. [1] is that it utilizes the recurrence relation of Gamma function and speeds up computation by truncating power-series.
2. Algorithm 1 is a correction of the method developed by Sra [2]. We note that the Algorithm in [2] for computing $I_s(x)$ omits the first term of power-series in (22) which leads to several orders of magnitude of errors. In addition, we discuss the evaluation of Gamma function $\Gamma(s+1)$.
3. The accuracy of the implementation of $I_s(x)$ depends on the Gamma function approximation and the convergence tolerance ε for truncating the power-series. Due to the efficient computation of the coefficients and easier error estimation, the Spouge formula should be chosen to obtain arbitrary order of accuracy [23].
4. Algorithm 1 is convergent for the cases $s \gg x$ or $s \sim x$, especially when the s is very large. Otherwise, one should resort to the more complicated method [26] or explicitly computing the modified Bessel function ratio [27].

5. Numerical experiments

In this section, we investigated extensive experiments to verify the superiority of our proposed methods. Firstly, the comparisons of different approximations with varying dimensionality or true κ[1] were showed in Table 1 and Figs. 1–4. Secondly, the experiments of implementing $I_s(x)$ were discussed in Tables 2 and 3. We conducted the experiments on a PC with Celeron 2.53 GHz CPU and 2GB RAM, running on Windows XP. We used MAPLE version 13 and set Digits := 1000.

[1] One can indeed run experiments with a true κ by solving (6), i.e. select a value of κ and then compute $A_d(\kappa)$ as accurately as possible. Then, use the resulting value of $\overline{R}$ in the nonlinear solver to estimate the $\hat{\kappa}$.

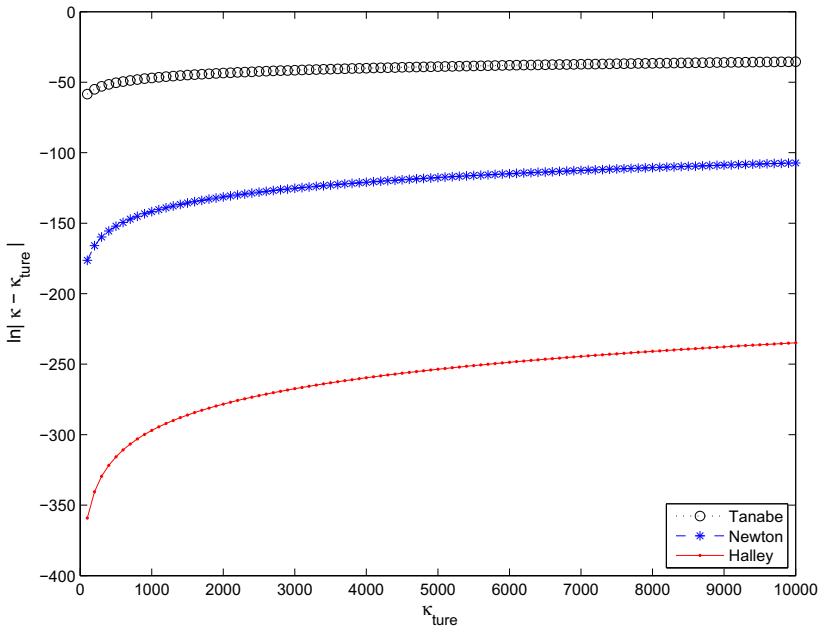

Fig. 2. Logarithmic absolute errors of approximations with varying κ from 100 to 10,000 and fixed d=1,000,000.

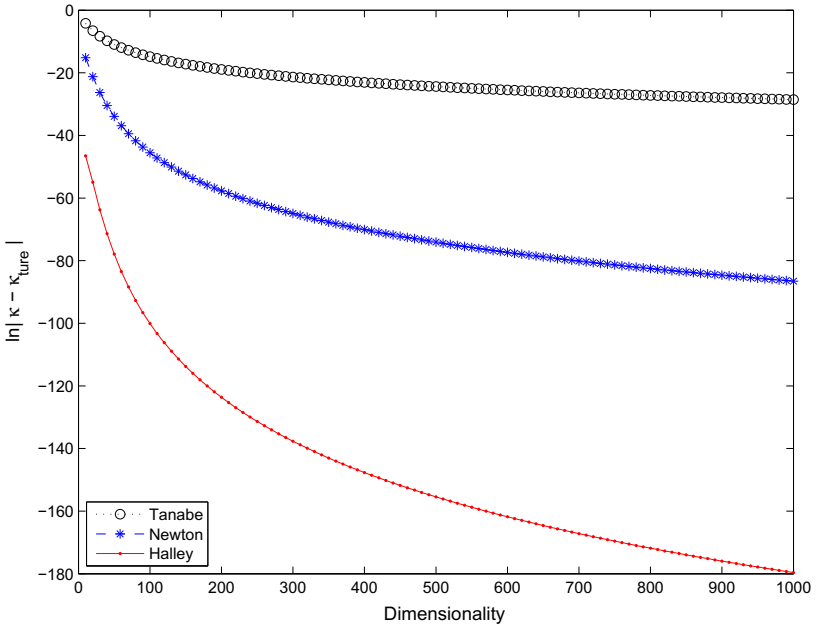

Fig. 3. Logarithmic absolute errors of approximations with varying d from 10 to 1000 and fixed κ=10.

5.1. Experiments for κ

In Table 1 we present four approximations given by (12), (11), (14) and (19) for different (d, κ_{true}) pairs. The approximation (19) outperforms other approximations (12), (11) and (14), and gains extremely high level of accuracy with increasing dimensionality of the data. From the table it is obvious that all the approximations become progressively better as dimensionality increases. Noticeable exceptions are the truncated Newton method (14) and truncated Halley method (19) for the case $d \sim \kappa_{true}$, and the errors give the highest level when $d = \kappa_{true}$. This is a consequence of $f'(\kappa) \approx 0$ when $d \sim \kappa_{true}$. The reason for the results lies in the fact that Householder's method has poor convergence properties near any point where the

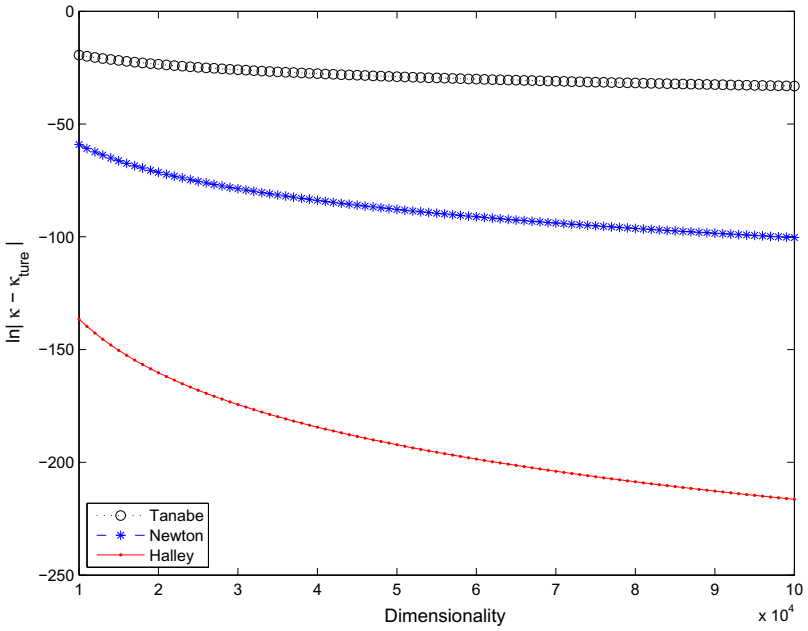

Fig. 4. Logarithmic absolute errors of approximations with varying d from 10,000 to 100,000 and fixed κ=1000.

Table 2
Number of iterations for truncating power-series with varying ε.

(s,x)	$\varepsilon=10^{-10}$	$\varepsilon=10^{-20}$	$\varepsilon=10^{-50}$	(s,x)	$\varepsilon=10^{-10}$	$\varepsilon=10^{-20}$	$\varepsilon=10^{-50}$
$(10^3,10^1)$	6	10	21	$(10^6,10^3)$	10	16	31
$(10^3,10^2)$	20	30	54	$(10^6,10^4)$	**64**	**86**	**133**
$(10^3,10^3)$	**296**	**341**	**432**	$(10^7,10^1)$	3	5	10
$(10^4,10^1)$	5	8	16	$(10^7,10^2)$	4	6	13
$(10^4,10^2)$	10	16	31	$(10^7,10^3)$	6	11	21
$(10^4,10^3)$	**64**	**85**	**132**	$(10^7,10^4)$	20	30	55
$(10^5,10^1)$	4	6	13	$(10^7,10^5)$	**356**	**410**	**522**
$(10^5,10^2)$	6	11	21	$(10^8,10^1)$	3	4	9
$(10^5,10^3)$	20	30	55	$(10^8,10^2)$	4	6	11
$(10^5,10^4)$	**355**	**409**	**520**	$(10^8,10^3)$	5	8	16
$(10^6,10^1)$	4	6	11	$(10^8,10^4)$	10	16	31
$(10^6,10^2)$	5	8	16	$(10^8,10^5)$	**64**	**86**	**133**

derivative $f'(\kappa)=0$ [4]. As shown in previous section, The approximation (11) proposed by Tanabe et al. [1] is not available for $d=2$.

Fig. 1, as well as Fig. 2, compares the logarithmic absolute errors of approximations for a fixed value of dimensionality d as κ_{true} is varied. From the figures, we observe that all the approximations become progressively worse as κ_{true} increases and the truncated Halley approximation (19) gives the smallest errors.

Next, Figs. 3 and 4 illustrate the comparisons of the logarithmic absolute errors of approximations as dimensionality d varies with fixed $\kappa_{true}=10,\ \kappa_{true}=1000$, respectively. From the figures, it is obvious that all the approximations become progressively better as d increases and the truncated Halley approximation (19) gives the smallest errors.

From these experiments, we conclude that the truncated Halley approximation (19) achieves the most accurate approximation in the parameter estimations for vMF distributions and that the significance becomes larger with increasing dimensionality of the data. Note that these experiments were conducted with 1000 digits of precision. The accuracy of approximation will, however, deteriorate using traditional double precision, which is why we consider higher-order approximations using the truncated Halley method.

5.2. Experiments for implementing $I_s(x)$

In this subsection, we illustrate the performance of the implementation of $I_s(x)$. We concentrate on high dimensional applications cases of $s \gg x$ or $s \sim x$, i.e. s is far greater than x or s is comparable in size to x , respectively.

Table 3
Running time (in seconds) of different methods for computing $I_s(x)$. All the results reported here are averages over 5 runs. A "–" indicates that the computation took too long to run.

(s,x)	Algorithm 1	MAPLE	Algorithm 1	MATHEMATICA
(1000,1000)	**0.028**	0.144	**0.000**	0.015
(1000,2000)	**0.028**	0.228	**0.015**	0.031
(1000,4000)	**0.047**	0.666	0.078	**0.016**
(2000,2000)	**0.019**	0.497	**0.016**	0.109
(2000,4000)	**0.041**	0.694	0.062	**0.047**
(2000,8000)	**0.119**	2.840	0.156	**0.016**
(4000,4000)	**0.025**	1.465	**0.047**	0.485
(4000,8000)	**0.075**	2.662	**0.125**	0.141
(4000,6000)	**0.372**	13.944	0.328	**0.031**
(8000,8000)	**0.056**	5.894	**0.094**	2.422
(8000,16,000)	**0.144**	11.897	**0.255**	0.653
(16,000,16,000)	**0.087**	26.787	**0.172**	13.203
(16,000,32,000)	**0.279**	62.078	**0.503**	3.248
(32,000,32,000)	**0.184**	150.766	**0.346**	68.536
(32,000,64,000)	**0.672**	348.500	**0.977**	17.828
(64,000,64,000)	**0.603**	837.719	**0.672**	395.538
(128,000,128,000)	**0.681**	4600.640	**1.344**	2005.453
(256,000,256,000)	**1.662**	–	**2.718**	–
(512,000,512,000)	**3.350**	–	**5.516**	–
(1,024,000,1,024,000)	**6.625**	–	**11.217**	–

Firstly, we investigate the number of iterations by truncating power-series varying convergence tolerance ε in Table 2 for high dimensional applications case $s \gg x$. We see that Algorithm 1 converges very quickly at most a few hundred iterations with a low tolerance. It is appealing to real world high dimensional applications.

Secondly, we compare the running time of Algorithm 1 that was implemented using MAPLE and MATHEMATICA, respectively, with the standard build-in functions. Table 3 shows the running time for the case $s \sim x$. From Table 3 we obtain the conclusion that our implementation, i.e., Algorithm 1, is faster than MAPLE and MATHEMATICA, most of time, especially in large magnitude.

6. Conclusions

In this paper, we discuss the parameter estimation for vMF distributions and present a high-order approximation using truncated Halley method. The comparison experiments show that our approach achieves significantly much more accurate approximation for the concentration parameter κ than the methods proposed by Tanabe et al. [1] and Sra [2]. In addition, we discuss the implementation of $I_s(x)$, the modified Bessel function of the first kind, which is the most time-consuming and important computation in vMF distributions related applications, such as the M-step of an Expectation Maximization (EM) algorithm based on mixture of vMF distributions [3]. Our implementation is faster than the build-in functions of MAPLE and MATHEMATICA, and is preferable for high dimensional applications. Our implementation also can be further improved by using C/C++ library for multiple-precision floating-point computations, e.g., MPFR [28].

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 60775055, 61074167). The first author thanks Suvrit Sra for his enlightened discussions.

References

[1] A. Tanabe, K. Fukumizu, S. Oba, T. Takenouchi, S. Ishii, Parameter estimation for von Mises–Fisher distributions, Comput. Stat. 22 (2007) 145–157.
[2] S. Sra, A short note on parameter approximation for von Mises–Fisher distributions: and a fast implementation of $i_s(x)$, Comput. Stat. 27 (1) (2011) 177–190.
[3] A. Banerjee, I.S. Dhillon, J. Ghosh, S. Sra, Clustering on the unit hypersphere using von Mises–Fisher distributions, J. Mach. Learn. Res. 6 (2005) 1345–1382.
[4] A.S. Householder, The Numerical Treatment of a Single Nonlinear Equation, McGraw-Hill, New York, USA, 1970.
[5] G. Watson, A Treatise on the Theory of Bessel Functions, second ed., Cambridge University Press, New York, USA, 1995.
[6] M. Abramowitz, I. Stegun, Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables, Courier Dover Publications, New York, USA, 1965.
[7] H. Schaeben, Normal orientation distributions, Texture Microstruct. 19 (1992) 197–202.
[8] G.S. Watson, E.J. Williams, On the construction of significance tests on the circle and the sphere, Biometrika 43 (3/4) (1956) 344–352.
[9] K.V. Mardia, P. Jupp, Directional Statistics, second ed., John Wiley and Sons Ltd., Chichester, UK, 2000.
[10] I. Dhillon, S. Sra, Modeling data using directional distributions, Technical Report, Department of Computer Sciences, University of Texas at Austin, TR-03-06, 2003.
[11] G. Ducharme, P. Milasevic, Estimating the concentration of the Langevin distribution, Can. J. Stat. 18 (2) (1990) 163–169.

[12] Y. Watamori, Statistical inference of Langevin distribution for directional data, Hiro. Math. J. 26 (1) (1996) 25–74.
[13] G. Hill, Evaluation and Inversion of the Ratios of Modified Bessel Functions, $I_1(x)/I_0(x)$ and $I_{1.5}(x)/I_{0.5}(x)$, ACM Trans. Math. Softw. 7 (2) (1981) 199–208.
[14] G. Schou, Estimation of the concentration parameter in von Mises–Fisher distributions, Biometrika 65 (2) (1978) 369–377.
[15] D. Dowe, J. Oliver, C. Wallace, MML estimation of the parameters of the spherical Fisher distribution, in: International Workshop on Algorithmic Learning Theory, Lecture Notes in Computer Science, vol. 1160, 1996, pp. 213–227.
[16] J. Traub, Iterative Methods for Solution of Equations, Prentice-Hall, Englewood NJ, USA, 1964.
[17] A. Melman, Geometry and convergence of Euler's and Halley's methods, SIAM Rev. 39 (4) (1997) 728–735.
[18] J. Kou, X. Wang, Sixth-order variants of Chebyshev-Halley methods for solving non-linear equations, Appl. Math. Comput. 190 (2) (2007) 1839–1843.
[19] W. Bi, Q. Wu, H. Ren, A new family of eighth-order iterative methods for solving nonlinear equations, Appl. Math. Comput. 214 (1) (2009) 236–245.
[20] L. Liu, X. Wang, Eighth-order methods with high efficiency index for solving nonlinear equations, Appl. Math. Comput. 215 (9) (2010) 3449–3454.
[21] G.R. Pugh, An analysis of the lanczos gamma approximation, Ph.D. Thesis, University of British Columbia, 2004.
[22] C. Lanczos, A precision approximation of the gamma function, SIAM J. Numer. Anal. 1 (1964) 86–96.
[23] J.L. Spouge, Computation of the gamma, digamma, and trigamma functions, SIAM J. Numer. Anal. 31 (3) (1994) 931–944.
[24] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, Cambridge University Press, New York, USA, 1992.
[25] A. Cuyt, V. Petersen, B. Verdonk, H. Waadeland, W. Jones, Handbook of Continued Fractions for Special Functions, Springer Verlag, Heidelberg, Germany, 2008.
[26] D. Amos, Computation of modified Bessel functions and their ratios, Math. Comput. 28 (125) (1974) 239–251.
[27] W. Gautschi, J. Slavik, On the computation of modified bessel function ratios, Math. Comput. 32 (143) (1978) 865–875.
[28] L. Fousse, G. Hanrot, V. Lefèvre, P. Pélissier, P. Zimmermann, MPFR: a multiple-precision binary floating-point library with correct rounding, ACM Trans. Math. Softw. 33 (2) (2007). Article 13.