MML Estimation of the Parameters of the Spherical Fisher Distribution

David L. Dowe, Jonathan J. Oliver and Chris S. Wallace

Department of Computer Science
Monash University, Clayton, Vic. 3168, Australia
e-mail: {dld, jono, csw}@cs.monash.edu.au

Abstract. The information-theoretic Minimum Message Length (MML) principle leads to a general invariant Bayesian technique for point estimation. We apply MML to the problem of estimating the concentration parameter, κ, of spherical Fisher distributions. (Assuming a uniform prior on the field direction, $\boldsymbol{\mu}$, MML simply returns the Maximum Likelihood estimate for $\boldsymbol{\mu}$.) In earlier work, we dealt with the von Mises circular case, $d = 2$. We say something about the general case for arbitrary $d \geq 2$ and how to derive the MML estimator, but here we only carry out a complete calculation for the spherical distribution, with $d = 3$. Our simulation results show that the MML estimator compares very favourably against the classical methods of Maximum Likelihood and marginal Maximum Likelihood (R.A. Fisher (1953), Schou (1978)). Our simulation results also show that the MML estimator compares quite favourably against alternative Bayesian methods.

1 Introduction

The spherical von Mises-Fisher distribution is a maximum entropy distribution of directions (unit vectors) on the surface of a unit sphere, being the expected long-term distribution of the direction of a unit dipole or pendulum constrained to the surface of the unit sphere and subjected to a uniform (magnetic or gravitational) field and random thermal fluctuations[1]. The mean of the distribution will be the direction of the field (or tendency), and the distribution will be symmetrical about its mean.

The general von Mises-Fisher distribution of arbitrary dimension [32] is of interest in a wide range of fields, such as cosmology [18], protein dihedral angles [4], biology, geography, geology, geophysics, medicine, meteorology and oceanography [10, 9]. The spherical Fisher distribution has been said to be "the most important distribution in directional data analysis" [22, Page 369], and has been the subject of some study [14, Chp. 8-9][15, 10].

The direction of the resultant vector of our observed data (the maximum likelihood estimate of the direction) can generally be seen to be the appropriate estimate for the direction of the tendency, although there are some exceptions to this. Two exceptions are if we are interested in the rejection of outliers or (as we

[1] A slightly longer version of this paper is available as a Technical Report[7].

are in subsequent work[20]) in modelling a mixture of more than one spherical Fisher distribution. Another exception will be if there is an informative (non-uniform) Bayesian prior on the direction. A final exception comes from using the Bayesian MAP estimate (which we discuss in Section 4.3 but which we do not advocate) under certain parameterisations.

In the work to follow, we provide a sketch of how to use the information-theoretic Message Length (MML)[31, 25, 26] principle to obtain the MML estimate for the concentration parameter, κ, of an arbitrary, d-dimensional, von Mises-Fisher distribution. (This generalises our earlier work[27, 28, 6] with $d = 2$.) We then obtain the MML estimate in the spherical Fisher case, with $d = 3$. We compare the MML estimator with alternative estimators, both Bayesian and classical, for this problem.

2 The von Mises-Fisher Distribution

The general, multi-dimensional von Mises-Fisher distribution was introduced by Watson and Williams [32] (following R.A. Fisher [11]), and estimating its parameters has been discussed in a number of contexts. Watson and Williams [32], Mardia [14, 15], Schou [22] and N.I. Fisher et al. [2, 10, 9] give estimators for its parameters in a classical framework.

Wallace and Dowe [27, 28] gave Minimum Message Length (MML) [25, 31] estimators for the circular (2-dimensional) von Mises-Fisher distribution, and Dowe, Oliver, Baxter and Wallace [6] compared MML with other Bayesian estimators for this circular case. In this paper, we extend these earlier MML works [27, 28, 6] to the spherical (3-dimensional) case.

The spherical von Mises-Fisher distribution corresponds to the distribution of the direction of a pendulum in a uniform gravitational field of direction, $\boldsymbol{\mu}$, with concentration parameter, κ. The concentration parameter, κ, can be thought of as the ratio of the field strength to the temperature of thermal fluctuations. For large κ, this closely approximates the 2-dimensional Gaussian distribution, $N(\boldsymbol{\mu}, \frac{1}{\kappa} I_2)$, where I_2 is the 2×2 identity matrix.

2.1 The Likelihood Function

Let $\mathbf{x}$ be a random vector on the surface of a d-dimensional sphere. The d-dimensional von Mises-Fisher distribution with mean vector $\boldsymbol{\mu}$ and concentration parameter, κ, has probability density function [22, Page 369]

$$f_d(\boldsymbol{\mu}, \kappa) = \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)} e^{\kappa\ \mathbf{x}.\boldsymbol{\mu}}$$

where $I_{\frac{d}{2}-1}(\kappa)$ is the modified Bessel function of the first kind of order $\frac{d}{2} - 1$, and $\mathbf{x}.\boldsymbol{\mu}$ is the dot (scalar) product of the vectors $\mathbf{x}$ and $\boldsymbol{\mu}$.

For the 3-dimensional case ($d = 3$) we find that[2]

$$I_{\frac{1}{2}}(\kappa) \;=\; \frac{\sqrt{2}\sinh(\kappa)}{\sqrt{\pi\kappa}}$$

We use the co-ordinate system from N.I.Fisher et al. [10, Page 18]. Thus, we represent the mean direction $\boldsymbol{\mu}$ as a co-latitude (α) and a longitude (β). Similarly we represent a data point, $\mathbf{x}$, as a co-latitude (θ) and a longitude (ϕ). The dot product of $\mathbf{x}$ and $\boldsymbol{\mu}$ is then:

$$\mathbf{x}.\boldsymbol{\mu} \;=\; \sin\theta\sin\alpha\cos(\phi-\beta) \;+\; \cos\theta\cos\alpha$$

and hence

$$f_3(\alpha,\ \beta,\ \kappa) = \frac{\kappa}{4\pi\sinh(\kappa)} e^{\kappa(\sin\theta\sin\alpha\cos(\phi-\beta)\;+\;\cos\theta\cos\alpha)}$$

For data $D = \{\mathbf{x(1)},\ \mathbf{x(2)},\ \ldots\ \mathbf{x(N)}\}$, the likelihood function is:

$$p(D|\alpha,\ \beta,\ \kappa) \;=\; \prod_{i=1}^{N} \frac{\kappa}{4\pi\sinh(\kappa)} e^{\kappa(\sin\theta_i\sin\alpha\cos(\phi_i-\beta)\;+\;\cos\theta_i\cos\alpha)}$$

and negative log-likelihood is:

$$\begin{aligned} L = & -N\log\kappa + N\log(4\pi\sinh(\kappa)) \\ & -\kappa\sum_{i=1}^{N}(\sin\theta_i\sin\alpha\cos(\phi_i-\beta) \;+\; \cos\theta_i\cos\alpha) \end{aligned}$$

3 Maximum Likelihood Estimator

3.1 The Maximum Likelihood Estimator for $\boldsymbol{\mu}$

The only term in the negative log-likelihood above which depends on $\boldsymbol{\mu}$ is the dot product of $\boldsymbol{\mu}$ with the sum of the $\mathbf{x(i)}$. Letting $\mathbf{R} \;=\; \sum_{i=1}^{N}\mathbf{x}(i)$ be the vector sum (resultant vector) of $\mathbf{x(1)}$, $\mathbf{x(2)}$, ..., $\mathbf{x(N)}$ and R be the length of $\mathbf{R}$, the term of the negative log-likelihood which depends on $\boldsymbol{\mu}$ is

$$-\kappa\ \mathbf{R}.\boldsymbol{\mu}$$

and hence the likelihood is maximised when

$$\boldsymbol{\mu} \;=\; \frac{\mathbf{R}}{R}$$

[2] This result can be confirmed by using Mathematica [33] to "Simplify[BesselI[1/2,k]]". Less opaquely, we show in the Appendix (between Acknowledgments and References) that this result can be obtained by integrating $e^{\kappa\mathbf{x}.\boldsymbol{\mu}}$ over the sphere by making the transformation with Jacobian $J = 1$ to the surface of a cylinder.

3.2 The Maximum Likelihood Estimator for κ

Given that we use the Maximum Likelihood estimate for μ in determining the Maximum Likelihood estimate for κ, we may re-write the negative log-likelihood as:

$$L(\kappa) = -N \log \kappa + N \log(4\pi) + N \log(\sinh(\kappa)) - \kappa R \tag{1}$$

To find the Maximum Likelihood estimate for κ, we differentiate L with respect to κ:

$$\frac{dL(\kappa)}{d\kappa} = \frac{-N}{\kappa} + N \coth \kappa - R$$

We numerically find κ_{MaxLik} by searching for the value of κ which sets $\frac{dL(\kappa)}{d\kappa} = 0$.

Having introduced the classical Maximum Likelihood estimator above for this problem, we now consider below some alternative Bayesian estimation techniques, including Minimum Message Length (MML) [25, 31].

4 Bayesian Estimation Techniques

Some fundamentalist Bayesians reject attempts to summarise a posterior density by an estimate as being unsound (e.g., Neal [16, Chp. 1]). For them, the posterior density is a sufficient and satisfactory end result of inference. However, we often wish to answer some question about the "real world", if possible, by summarising the posterior using an estimate. In a traditional Bayesian framework, a single estimate may be found if there is a clearly defined and known loss function. This function specifies the "loss" occurring if the estimate is used when the estimate is not the true value.

In this section, we argue why we sometimes prefer to perform inference without a loss function. We go on to consider two Bayesian methods for point estimation:

- the mode of the posterior density (MAP) [1, Page 257] and
- the Minimum Message Length (MML) estimate [25, 31].

Both of these estimators are Bayesian and require a prior distribution.

4.1 Loss Functions

It is not always the case that a loss function is available for applications. Let us consider the question:

> Does the universe have a preferred direction?

One method of attacking this problem is to look at the distribution of micro-wave background radiation (or of galaxies) in the universe. If the galaxies are arranged according to a spherical von Mises-Fisher distribution with a non-zero κ, then we may be willing to assert that the universe did have a preferred direction. We can not see how one would determine a 'preferred' loss function for this problem.

In the rest of this section, we consider Bayesian point estimation using the posterior distribution without a loss function. We consider the Bayesian MAP estimate and the (invariant) MML estimate.

4.2 Prior Distributions

We assume a prior[3] which is uniform in direction [10, Page 84] and independent of κ:

$$h_{\alpha,\,\beta}(\alpha,\,\beta) \;=\; \frac{\sin(\alpha)}{4\pi}$$

We consider a generalisation of a prior on κ used sucessfully for the 2-dimensional von Mises distribution [27, 28, 6]:

$$h_{\kappa}(\kappa) \propto \frac{\kappa^{d-1}}{(1+\kappa^2)^{\frac{d+1}{2}}}$$

which, in 3 dimensions, is

$$h_{\kappa}(\kappa) = \frac{4\ \kappa^2}{\pi\ (1+\kappa^2)^2}$$

and hence:

$$h_{\alpha,\,\beta,\,\kappa}(\alpha,\,\beta,\,\kappa) \;=\; h_{\alpha,\,\beta}(\alpha,\,\beta) \times h_{\kappa}(\kappa) \;=\; \frac{\kappa^2\ \sin(\alpha)}{\pi^2\ (1+\kappa^2)^2} \tag{2}$$

This prior is indeed uniform in direction, since if we transform this prior to Cartesian co-ordinates[4] we get

$$h_{x,\,y,\,z}(x,\,y,\,z) \;=\; \frac{1}{\pi^2\ (1+x^2+y^2+z^2)^2}$$

4.3 The MAP Estimate

The MAP estimate is the value of the parameters which maximises the posterior density function [1, Page 257] [3, Page 64]. It is known to be, in general, not invariant under re-parameterisation. In the 2-dimensional von Mises case, Oliver and Baxter [19] demonstrated how the MAP estimate re-locates with the simple re-parameterisation of changing from polar to Cartesian co-ordinates. Let us now briefly examine this issue of dependence of the MAP estimate on the co-ordinate system for the 3-dimensional, spherical, case.

[3] Where genuine prior information is available, we advocate its best mathematical articulation. We object to a Jeffreys prior[12] on the grounds that the expected Fisher information will be a function of the measuring apparatus, and so the choice of a Jeffreys prior, while mathematically convenient, would suggest the possession of a genuine Bayesian prior belief that the properties of the world that one wishes to study are quite strongly dependent upon one's (possibly arbitrary) choice of measuring apparatus.

The priors we choose here are quite "colourless" - uniform in direction, normalisable and locally uniform at the Cartesian origin in κ.

[4] We transform a prior by dividing by the Jacobian of the transformation, which is $\kappa^2\ \sin(\alpha)$ [23, Chp. 7]. Our transformation is $x = \kappa\cos(\beta)\sin(\alpha)$, $y = \kappa\sin(\beta)\sin(\alpha)$, $z = \kappa\cos(\alpha)$.

The MAP estimate (in Spherical Co-ordinates) is the value of $(\alpha,\ \beta,\ \kappa)$ which maximises the expression:

$$h_{\alpha,\ \beta,\ \kappa}(\alpha,\ \beta,\ \kappa)\ \times\ p(D|\alpha,\ \beta,\ \kappa) \tag{3}$$

where $h_{\alpha,\ \beta,\ \kappa}(\alpha,\ \beta,\ \kappa)$ is our prior distribution over spherical parameter values.

We might like to compare the above MAP estimate with the MAP estimate in Cartesian Co-ordinates. The MAP estimate (in Cartesian Co-ordinates) is the value of $x,\ y,\ z$ which maximises:

$$h_{x,\ y,\ z}(x,\ y,\ z)\ \times\ p(D|x,\ y,\ z) \tag{4}$$

Again, this is a different vector in general than the MAP estimate in Spherical Co-ordinates. One way of seeing this clearly is as follows:

Let the (spherical) parameterisation of the field be $\boldsymbol{\theta}$. Then the MAP estimate in that parameterisation will be an equivalent vector to $(x',\ y',\ z')$, where $x',\ y',\ z'$ are the values which maximise:

$$J\ \times\ h_{x,\ y,\ z}(x',\ y',\ z')\ \times\ p(D|x',\ y',\ z')$$

where J is the Jacobian of the transformation between $\boldsymbol{\theta}$ and Cartesian Co-ordinates.

It is unclear which parameterisation is 'best' (or "natural"[21]) for estimation. Furthermore, we also have to select other features for some of these parameterisations. For example, in Cylindrical Co-ordinates we must choose which direction we will align the z-axis with. Such a choice affects the Bayesian MAP estimate (which we do not advocate, as we discuss below). We will see shortly that the MML estimate remains invariant.

4.3.1 A Problem with the MAP Estimate in Spherical Co-ordinates

It would appear that the "natural" parameterisation for the spherical von Mises-Fisher distribution would be Spherical Co-ordinates, since the likelihood function is expressed in these terms, and the parameter of most interest is the strength of the field.

Consider the situation when we wish to determine the MAP estimate in spherical co-ordinates, $(\alpha,\ \beta,\ \kappa)$, for some data using the prior in Equation (2). Before we collect the data, we establish an origin for our spherical co-ordinates. This involves selecting a direction where $\alpha\ =\ \frac{\pi}{2}$ (i.e., the equator, where the co-latitude is 90 degrees), and a direction where $\beta\ =\ 0$ (i.e., where the longitude is zero, the Greenwich meridian). The choice of where we defined our equator affects the MAP estimate. Consider the situation when the data is generated from a field which is aligned through the North pole (or the South pole). Equation (2) assigns 0 prior to this field (since $\sin(\alpha)\ =\ 0$). The MAP estimate will never return the correct field and will be strongly discouraged from returning a nearby field. (This result will, incidentally, get proportionally worse as $\sin^{d-2}$ as we increase the dimensionality, d.)

To resolve this problem, we define the MAP estimate in spherical co-ordinates to be the MAP estimate when we align the point ($\alpha = \frac{\pi}{2}$, $\beta = 0$) with the Maximum Likelihood estimate of the direction of the field. Difficulty in choosing a "natural" parameterisation can present potential problems for the Bayesian MAP method and for other methods [21] which are not invariant under reparameterisation.

4.4 Invariant Bayesian Estimation Methods

Many parameter estimation problems do not come accompanied by associated loss functions, and we believe that parameterisation should not affect point estimation. The Minimum Message Length (MML) [25, 26, 31] and the Minimum Expected Kullback-Leibler distance[5] estimators are two invariant Bayesian point estimation techniques. We believe that this invariance is a very desirable property, especially when there are [23, Pages 137-140][10, Pages 17-22] some 10 or more parameterisations of the sphere in relatively common usage, with spherical, Cartesian and cylindrical co-ordinates being but three.

4.5 The Minimum Message Length (MML) Estimates

MML is an invariant Bayesian point estimation method proposed by Wallace et al. [25, 26, 31]. The underlying idea is to maximise the posterior probability of an hypothesis, H, given data, D, but the approach taken is somewhat different to that in Bayesian MAP estimation. A list of MML applications and an elaboration on how MML maximises the posterior probability are given in [29, Page 37]. Further introductory MML material is given in [27, 19], with the invariance of MML being discussed in [26], [31, p245], [27, Pages 1-3] and [19, Section 5.4].

4.5.1 The Message Length Formula

The MML estimate is the value of (α, β, κ) which minimises the message length expression [31, Page 245]:

$$MessLen(\alpha,\ \beta,\ \kappa\ \&\ D) = -log(\frac{h_{\alpha,\ \beta,\ \kappa}(\alpha,\ \beta,\ \kappa)\ f_3(D|\alpha,\ \beta,\ \kappa)}{\sqrt{det(F(\alpha,\ \beta,\ \kappa))}}) + \text{Constants},$$

where $det(F(\alpha,\ \beta,\ \kappa))$ is the determinant of the expected Fisher Information Matrix. The likelihood and log-likelihood were discussed in Section 2.1, and the prior – uniform in direction and locally uniform near the Cartesian origin – is discussed in Section 4.2.

We interpret the term $\frac{1}{\sqrt{det(F(\alpha,\ \beta,\ \kappa))}}$ as being proportional to the volume of uncertainty that we have in our MML estimates. Minimising the Message Length is then equivalent to maximising:

$$\text{Exp} = \frac{h_{\alpha,\ \beta,\ \kappa}(\alpha,\ \beta,\ \kappa)\ f_3(D|\alpha,\ \beta,\ \kappa)}{\sqrt{det(F(\alpha,\ \beta,\ \kappa))}} \tag{5}$$