Modeling Data using Directional Distributions

Inderjit S. Dhillon and Suvrit Sra
Department of Computer Sciences
The University of Texas at Austin
Austin, TX 78712
{suvrit,inderjit}@cs.utexas.edu

25 January, 2003*

Technical Report # TR-03-06

Abstract

Traditionally multi-variate normal distributions have been the staple of data modeling in most domains. For some domains, the model they provide is either inadequate or incorrect because of the disregard for the directional components of the data. We present a generative model for data that is suitable for modeling directional data (as can arise in text and gene expression clustering). We use mixtures of *von Mises-Fisher* distributions to model our data since the von Mises-Fisher distribution is the *natural* distribution for directional data. We derive an Expectation Maximization (EM) algorithm to find the maximum likelihood estimates for the parameters of our mixture model, and provide various experimental results to evaluate the "correctness" of our formulation. In this paper we also provide some of the mathematical background necessary to carry out all the derivations and to gain insight for an implementation.

1 Introduction

Traditional statistical approaches involve multi-variate data drawn from $\mathbb{R}^p$, and little or no significance is attached to the directional nature of the observed data. For many phenomena or processes it makes more sense to consider the directional components of the data involved, rather than just the magnitude alone. For example, modeling wind current directions, modeling geomagnetism, and measurements derived from clocks and compasses all seem to require a directional model [MJ00]. A much wider array of fields and contexts in which directional data arises is enlisted in [MJ00], and the interested reader is urged to atleast gloss over that information.

A fundamental distribution on the circle called the *von Mises* distribution was first introduced by von Mises [vM18]. We address the issue of modeling data using the von Mises-Fisher (vMF) distribution [MJ00], which is a generalization (to higher dimensions) of the von Mises distribution. We concentrate on using the vMF distribution as it is a distribution that arises naturally for directional data—akin to the multivariate Normal distribution ([MJ00, pp. 171-172]). Furthermore, it has been observed that in high dimensional text data, cosine similarity performs much better than a Euclidean distance metric[1] [DFG01]. This observation suggests following a directional model for the text data rather than ascribing significance to a magnitude based (or traditional) model.

[*] Revised 7th June, 2003.

[1] Empirically cosine similarity has been observed to outperform Euclidean or Mahalanobis type distance measures in information retrieval tasks.

Another application domain for the vMF model is modeling gene micro-array data (gene expression data). Gene expression data has been found to have unique directional characteristics that suggest the use of a directional model for modeling it. Recently Dhillon et. al (see [DMR03]) have found that gene expression data yields interesting diametric clusters. Intuitively these clusters could be thought of as data pointing in opposite directions, hinting at the underlying importance of directional orientation[2].

For text data, one byproduct of using a generative model like a mixture of vMF distributions, is the ability to obtain a soft-clustering of the data. The need for soft-clustering comes to the foreground when the text collections to be clustered can have documents with multiple labels. A more accurate generative model can also serve as an aid for improved classification for text data, especially where more meaningful soft labels are desired[3].

Organization of this report

The remainder of this report is organized as follows. Section 2 presents the multi-variate von Mises-Fisher distribution. Section 3 carries out the maximum likelihood estimation of parameters for data drawn from a single vMF distribution. Section 4 derives and presents the EM algorithm for estimating parameters for data drawn from a mixture of vMFs. In section 5 we show the results of experimentation with simulated mixtures of vMF distributions. Section 6 concludes this report. Some useful mathematical details are furnished by Appendices A and B. Appendix A provides mathematical background that is useful in general for understanding the derivations and Appendix B offers a brief primer on directional distributions.

2 The von Mises-Fisher Distribution

A p-dimensional unit random vector $\mathbf{x}$ ($\|\mathbf{x}\| = 1$) is said to have p-variate *von Mises-Fisher* distribution $M_p(\boldsymbol{\mu}, \kappa)$ if its probability density is:

$$c_p(\kappa) e^{\kappa \boldsymbol{\mu}^T \mathbf{x}}, \quad \mathbf{x} \in S^{p-1}, \tag{2.1}$$

where $\|\boldsymbol{\mu}\| = 1$, $\kappa \geq 0$, S^{p-1} is the p dimensional unit hypersphere (also denoted as S_p in some literature), and $c_p(\kappa)$ the normalizing constant is given by (see B.2)

$$c_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)}. \tag{2.2}$$

For more details the interested reader is urged to look at Appendix B.

2.1 Example vMF distribution

In two dimensions (on the circle S^0), the probability density assumes the form

$$g(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \quad 0 \leq \theta < 2\pi. \tag{2.3}$$

This is called the *von Mises* distribution. Figure 1 shows a plot of this density with mean at 0 radians and for $\kappa \in \{0, 0.3, 1, 4, 20\}$.

From the figure we can see that as κ increases the density becomes more and more concentrated about the mean direction. Thus κ is called the concentration parameter.

[2] Most clustering algorithms for gene expression data use Pearson correlation, which equals cosine similarity of transformed vectors, and thus our directional model should fit it well.

[3] Though given the nature of high-dimensional sparse data and models based on some member of the exponential family of distributions, the ability to obtain useful soft-label remains difficult without explicitly imposing "softness" constraints.

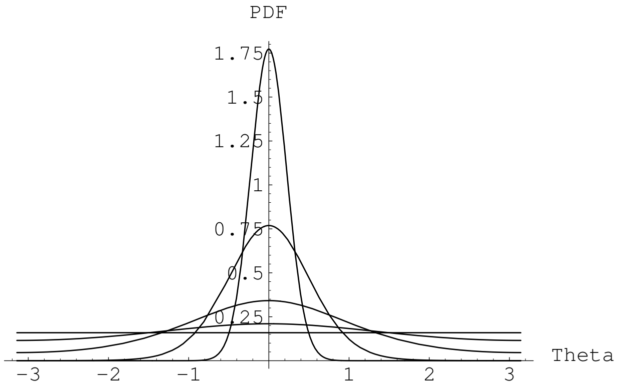


Figure 1: von Mises distribution for various κ ($\kappa = 0, 0.3, 1, 4, 20$).

3 Maximum Likelihood Estimates for von Mises-Fisher Distributions

Let $\mathscr{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be the set of sample unit vectors following $M_p(\boldsymbol{\mu}, \kappa)$. Since each $\mathbf{x}_i \in \mathscr{D}$ is assumed to be independent the likelihood is

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n | \kappa, \boldsymbol{\mu}) = \prod_{i=1}^{n} c_p(\kappa) e^{\kappa \boldsymbol{\mu}^T \mathbf{x}_i}. \tag{3.1}$$

Thus the log-likelihood is

$$\mathscr{L}(\kappa, \boldsymbol{\mu}) = n \ln c_p(\kappa) + \kappa \boldsymbol{\mu}^T \mathbf{r}, \tag{3.2}$$

where $\mathbf{r} = \sum_i \mathbf{x}_i$ is the resultant vector. Differentiating (3.2) w.r.t $\boldsymbol{\mu}$ subject to the constraint $\boldsymbol{\mu}^T \boldsymbol{\mu} = 1$ we get

$$\kappa \mathbf{r} = 2\lambda \boldsymbol{\mu}, \tag{3.3}$$

where λ is a Lagrange multiplier. Let $\hat{\boldsymbol{\mu}}$ denote the m.l.e. for $\boldsymbol{\mu}$. From (3.3) and the fact that $\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}} = 1$ we conclude that $\hat{\boldsymbol{\mu}} = \mathbf{r}/\|\mathbf{r}\|$. Let us write $\|\mathbf{r}\| = n\bar{R}$, where $\bar{R}$ denotes the average resultant length.

Differentiating (3.2) w.r.t κ we obtain

$$\frac{n c_p'(\kappa)}{c_p(\kappa)} + n\bar{R} = 0. \tag{3.4}$$

For brevity, let us write $s = p/2 - 1$. From (2.1),

$$c_p'(\kappa) = \frac{s\kappa^{s-1}}{\alpha I_s(\kappa)} - \frac{\kappa^s I_s'(\kappa)}{\alpha I_s^2(\kappa)}, \tag{3.5}$$

where $\alpha = (2\pi)^{s+1}$ is a constant. We thus simplify $c_p'(\kappa)/c_p(\kappa)$ to be

$$\frac{s}{\kappa} - \frac{I_s'(\kappa)}{I_s(\kappa)}. \tag{3.6}$$

Using the fact that (see for e.g., [AS74])

$$\kappa I_{s+1}(\kappa) = \kappa I_s'(\kappa) - s I_s(\kappa), \tag{3.7}$$

we obtain

$$\frac{-c_p'(\kappa)}{c_p(\kappa)} = \frac{I_{s+1}(\kappa)}{I_s(\kappa)} = \frac{I_{p/2}(\kappa)}{I_{p/2-1}(\kappa)} = A_p(\kappa). \tag{3.8}$$

Using (3.4) and (3.8) we find that the m.l.e. for κ is given by

$$\hat{\kappa} = A_p^{-1}(\bar{R}). \tag{3.9}$$

Since $A_p(\kappa)$ is the ratio of Bessel functions we cannot obtain a closed form functional inverse. Hence to solve for $A_p^{-1}(\bar{R})$ we have to resort to numerical or asymptotic methods.

For large values of κ the following approximation is well known ([AS74], Chapter 9):

$$I_p(\kappa) \approx \frac{1}{\sqrt{2\pi\kappa}} e^{\kappa} \left(1 - \frac{4p^2 - 1}{8\kappa}\right). \tag{3.10}$$

Using (3.10) we obtain

$$A_p(\kappa) \approx \left(1 - \frac{p^2 - 1}{8\kappa}\right) \left(1 - \frac{(p-2)^2 - 1}{8\kappa}\right)^{-1}. \tag{3.11}$$

Now using the fact that κ is large, expanding the second term using the binomial theorem and ignoring terms that have squares or higher powers of κ in the denominator we are left with

$$A_p(\kappa) \approx \left(1 - \frac{p^2 - 1}{8\kappa}\right) \left(1 + \frac{(p-2)^2 - 1}{8\kappa}\right). \tag{3.12}$$

On again ignoring terms containing κ^2 in the denominator we finally have

$$A_p(\kappa) \approx 1 - \frac{p-1}{2\kappa}. \tag{3.13}$$

Hence for large κ we obtain

$$\hat{\kappa} = \frac{\frac{1}{2}(p-1)}{1 - \bar{R}}. \tag{3.14}$$

We can write $I_p(\kappa)$ as (A.8),

$$I_p(\kappa) = \sum_{k \geq 0} \frac{1}{\Gamma(k+p+1) k!} \left(\frac{\kappa}{2}\right)^{2k+p}. \tag{3.15}$$

For small κ we use only the first two terms of this series, ignoring terms with higher powers of κ to get

$$I_p(\kappa) \approx \frac{\kappa^p}{2^p\, p!} + \frac{\kappa^{2+p}}{2^{p+2}\,(1+p)!}. \tag{3.16}$$

Using (3.16) and on simplifying $A_p(\kappa)$ we obtain

$$A_p(\kappa) \approx \frac{\kappa}{p}, \tag{3.17}$$

so that,

$$\hat{\kappa} = p\bar{R}. \tag{3.18}$$

See [MJ00] for conditions under which the approximations for $\hat{\kappa}$ are valid, at least for $p = 2, 3$.

These approximations for κ do not really take into account the dimensionality of the data and thus for high dimensions (when κ is big by itself but κ/p is not very small or very big) these estimates fail to yield sufficient accuracy. We have found that the following seems to yield a very reasonable approximation most of the time (Appendix B gives a derivation):

$$\hat{\kappa} = \frac{\bar{R}p - \bar{R}^3}{1 - \bar{R}^2}. \tag{3.19}$$

While implementing the calculation of $A_p(\kappa)$ on a computer it pays to implement it as a continued fraction. To solve for κ we can use the approximation given by (3.19) as a starting point and then do a couple of Newton-Raphson iterations to improve our guess, though most often we do not really need very accurate approximations of κ and (3.19) suffices. Some further details can be found in Appendix B.

4 EM for a vMF mixture

In this section we derive the mixture-density parameter update equations for a mixture of von Mises-Fisher distributions. First we obtain the maximum likelihood estimates (m.l.e.) assuming complete data and then adapt to the incomplete data case viewing the problem in an Expectation Maximization (EM) framework. The Maximum Likelihood Estimates are derived using the method given in §10.3 of [DHS00].

4.1 Maximum Likelihood Estimates

Suppose that we are given a set $\mathscr{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ of n unlabeled samples drawn independently from the mixture density:

$$p(\mathbf{x}|\Theta) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \theta_j)P(\omega_j), \tag{4.1}$$

where $\omega_1, \ldots, \omega_c$ are the c classes from which data can come. The full parameter vector Θ is fixed but unknown. Since the $\mathbf{x}_i$ are assumed to be independent the likelihood can be written as

$$p(\mathscr{D}|\Theta) = \prod_{k=1}^{n} p(\mathbf{x}_k|\Theta). \tag{4.2}$$

The M.L.E. $\hat{\Theta}$ is that value of Θ that maximizes $p(\mathscr{D}|\Theta)$. Now let $\mathscr{L}(\mathscr{D}|\Theta)$ be the log-likelihood given by

$$\mathscr{L}(\mathscr{D}|\Theta) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\Theta). \tag{4.3}$$

Note that $\Theta = (\theta_1, \ldots, \theta_c)^T$ is the total parameter vector; θ_i is the parameter vector for class i. We can write the gradient of the log-likelihood w.r.t. θ_i as,

$$\nabla_{\theta_i}\mathscr{L}(\Theta) = \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\Theta)} \nabla_{\theta_i} \left(\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j, \theta_j)P(\omega_j) \right). \tag{4.4}$$

Now if we assume that θ_i is functionally independent of θ_j then we can simplify the above equations. First let us introduce the posterior probability (using Bayes' rule):

$$P(\omega_i|\mathbf{x}_k, \Theta) = \frac{p(\omega_i, \mathbf{x}_k|\Theta)}{p(\mathbf{x}_k|\Theta)} = \frac{p(\mathbf{x}_k|\omega_i, \Theta)P(\omega_i)}{p(\mathbf{x}_k|\Theta)}. \tag{4.5}$$

Using this posterior probability we see that the m.l.e. $\hat{\theta}_i$ must satisfy:

$$\sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \Theta)\nabla_{\theta_i} \ln p(\mathbf{x}_k|\omega_i, \hat{\theta}_i) = 0 \qquad i = 1, \ldots, c. \tag{4.6}$$

If the priors are also unknown then maximizing (4.6), subject to the condition

$$\sum_{j=1}^{c} \hat{P}(\omega_j) = 1, \tag{4.7}$$

we obtain the following m.l.e. for them:

$$\hat{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\Theta}). \tag{4.8}$$

4.1.1 Maximum Likelihood for vMF mixture

The derivation given in the previous section applies to any probability density. Assuming that each sample $\mathbf{x}_k$ comes from a von Mises-Fisher distribution, i.e.

$$p(\mathbf{x}_k|\omega_i, \theta_i) = c_p(\kappa_i)e^{\kappa_i \boldsymbol{\mu}_i^T \mathbf{x}_k}, \tag{4.9}$$

we can solve the above maximum likelihood equations to obtain values of the parameters $(\kappa_i, \boldsymbol{\mu}_i)$ for $i = 1, \ldots, c$. We maximize the log-likelihood w.r.t. $\boldsymbol{\mu}_i$ subject to the constraint $\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i = 1$ to obtain:

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\Theta})\mathbf{x}_k}{\|\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\Theta})\mathbf{x}_k\|}. \tag{4.10}$$

Writing $-c_p'(\kappa_i)/c_p(\kappa_i) = A_p(\kappa_i)$ as usual, we obtain the following m.l.e. equation for κ_i:

$$A_p(\hat{\kappa}_i) = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\Theta})\hat{\boldsymbol{\mu}}_i^T \mathbf{x}_k}{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\Theta})}. \tag{4.11}$$

Hence we obtain $\hat{\kappa}_i$ by calculating $A_p^{-1}(\cdot)$ for the above argument (see Section 3).

In all these equations the value of the posterior probability is given by:

$$\hat{P}(\omega_i|\mathbf{x}_k, \hat{\Theta}) = \frac{c_p(\kappa_i)e^{\kappa_i \boldsymbol{\mu}_i^T \mathbf{x}_k}\hat{P}(\omega_i)}{\sum_{j=1}^{c} c_p(\kappa_j)e^{\kappa_j \boldsymbol{\mu}_j^T \mathbf{x}_k}\hat{P}(\omega_j)}. \tag{4.12}$$

From these equations it seems that the posterior probability is large when: $c_p(\kappa_i)$ is large and when $\kappa_i \boldsymbol{\mu}_i^T \mathbf{x}_k$ is large. We could thus use these in an explicit objective function while iteratively calculating the m.l.e. for the parameters.

4.2 Parameter estimation using EM

For unlabeled data points the class to which a given data point belongs is not known. In the presence of such incomplete data we have to take resort to an Expectation Maximization scheme for calculating the m.l.e. for parameters. We have the following probabilistic model

$$p(\mathbf{x}|\Theta) = \sum_{j=1}^{c} \alpha_j p(\mathbf{x}|\theta_j), \tag{4.13}$$

where the α_j's are the so called "mixing" parameters (or class priors) and Θ is the parameter vector for the mixture model. The incomplete-data log-likelihood expression for this density from the data $\mathscr{D}$ is given by:

$$\mathscr{L}(\mathscr{D}|\Theta) = \ln \prod_{k=1}^{n} p(\mathbf{x}_k|\Theta). \tag{4.14}$$

Now if we consider $\mathscr{D}$ to be incomplete, but assume the existence of unobserved data items $\mathscr{Y} = \{y_i\}_{i=1}^{n}$, whose values inform us which component density generated each data item, the problem becomes easier. That is to say that each y_i here corresponds to some ω_j as discussed in the previous section. We let, $y_i = k$ if the i^{th} sample $\mathbf{x}_i$ was generated by the mixture component corresponding to ω_k. Thus we can look at the m.l.e. derivation in the previous section in a manner similar to the one given by [Bil97]. After some tedious algebra we essentially reach the same equations as given in the previous section. The scheme to perform the calculations is an EM algorithm that proceeds by iterative updates to estimate the parameters of the assumed distribution on data.

Algorithm Estimate $\alpha_j, \boldsymbol{\mu}_j, \kappa_j$ for $1 \leq j \leq c$

```
0: Initialize all $\alpha_j, \boldsymbol{\mu}_j, \kappa_j, P(\omega_j|\mathbf{x}_k, \theta)$
2. repeat
3.     for $k = 1$ to $N$ do
4.         for $j = 1$ to $c$ do
5.             $p(\mathbf{x}_k|\omega_j, \Theta) = c_p(\kappa_j) e^{\kappa_j \boldsymbol{\mu}_j^T \mathbf{x}_k}$
               $\hat{P}(\omega_j|\mathbf{x}_k, \hat{\Theta}) = \frac{p(\mathbf{x}_k|\omega_j, \Theta)\alpha_j}{\sum_{l=1}^{c} p(\mathbf{x}_k|\omega_l, \Theta)\alpha_l}$
6.         end
7.     end
8.     for $j = 1$ to $c$ do
9.         $n_j = \sum_{k=1}^{n} \hat{P}(\omega_j|\mathbf{x}_k, \hat{\Theta})$
           $\hat{\alpha}_j = n_j/n$
           $\mathbf{r}_j = \sum_{k=1}^{n} \hat{P}(\omega_j|\mathbf{x}_k, \hat{\Theta})\mathbf{x}_k$
           $\hat{\boldsymbol{\mu}}_j = \mathbf{r}_j/\|\mathbf{r}_j\|$
           $\hat{\kappa}_j = A_p^{-1}(\|\mathbf{r}_j\|/n_j)$
10.    end
11. until stopping criteria met.
```

Figure 2: EM algorithm for a mixture of vMF distributions.

4.3 Implementation Details

The above algorithm was implemented in MATLAB and its source is available upon request. The calculation of $\hat{\kappa}_j$ in step 9 above is implemented using the approximation given by (3.19).

There are various ways in which we could initialize our EM algorithm. An easy and effective method is to initialize the original guesses of the mean directions by using a spherical k-means type algorithm [DM01], and calculate the initial values of the parameters from the clustering obtained.

5 Experiments

This section discusses some of the experiments performed and the results obtained. We tested our algorithm on data sampled from simulated mixtures of vMFs.

5.1 Simulation of vMF mixtures

This information is adapted from Chapter 10 of [MJ00]. For $\kappa > 0$, the associated vMF distribution has a mode at the mean direction $\boldsymbol{\mu}$, whereas when $\kappa = 0$ the distribution is uniform. The larger the value of κ, the greater is the clustering around the mean direction.

Since the vMF density depends on $\mathbf{x}$ only through $\boldsymbol{\mu}^T\mathbf{x}$, this distribution is rotationally symmetric about $\boldsymbol{\mu}$. Further in the tangent normal decomposition:

$$\mathbf{x} = t\boldsymbol{\mu} + (1-t^2)^{1/2}\boldsymbol{\zeta}, \tag{5.1}$$

t is invariant under rotation about $\boldsymbol{\mu}$ while $\boldsymbol{\zeta}$ is equivariant (i.e. any such rotation Q takes $\boldsymbol{\zeta}$ to $Q\boldsymbol{\zeta}$). Thus the conditional distribution $\boldsymbol{\zeta}|t$ is uniform on S^{p-2}. It follows that $\boldsymbol{\zeta}$ and t are independent and $\boldsymbol{\zeta}$ is uniform on S^{p-2}. Further (see [MJ00]), we see that the marginal density of t is:

$$\frac{\left(\frac{\kappa}{2}\right)^{\frac{p}{2}-1}}{\Gamma(\frac{p-1}{2})\Gamma(\frac{1}{2})I_{\frac{p-1}{2}}(\kappa)} e^{\kappa t}(1-t^2)^{\frac{p-3}{2}}, \tag{5.2}$$

on the interval $[-1, 1]$.

```
function mixsamp($n$, $d$, $M$)
In: $n$ points to sample; $d$ dimensionality, $M$ mixture data structure
Out: $M$ modified mixture, $L$ label of each sampled point.

$L \leftarrow$ zeros(n,1);
$P \leftarrow$ rand(1,n);
$\mathscr{X} \leftarrow$ zeros(n,d);
$cp \leftarrow 0$;            {Cumulative sum of priors}
$cs \leftarrow 0$;            {Cumulative sum of number of sampled points}
for $j \leftarrow 1$ to $k$
    $ns \leftarrow$ sum($P \geq cp$ and $P < cp + M.P(\omega_j)$);
    $\kappa \leftarrow M.\kappa(j)$;
    $\mathscr{X}(ns+1 : cs+ns, :) \leftarrow$ vsamp($M.\boldsymbol{\mu}_j$, $\kappa$, $ns$);
    $L(cs+1 : cs+ns) \leftarrow j$;
    $cp \leftarrow cp + M.P(\omega_j)$;
    $cs \leftarrow cs + ns$;
end
$M.\mathscr{X} \leftarrow \mathscr{X}$
```

Figure 3: Simulating a mixture of vMFs.

From the facts that $\boldsymbol{\zeta}$ and t are independent and that $\boldsymbol{\zeta}$ is uniformly distributed on S^{p-2} it follows that the simulation of a vMF is easy. If $\boldsymbol{\zeta}$ and t are generated independently from the uniform distribution on S^{p-2} and from (5.2) respectively then

$$\mathbf{x} = t\boldsymbol{\mu} + (1-t^2)^{1/2}\boldsymbol{\zeta},$$

is a pseudo-random unit vector with the $M_p(\boldsymbol{\mu}, \kappa)$ distribution. Further information about this can be found in [Woo94]. We used the MATLAB Statistics Toolbox for aiding our implementation of Wood's algorithm ([Woo94]). Figure 4 gives Wood's algorithm (slight adaptation) that we used to simulate a single vMF distribution. Figure 3 gives the algorithm used to simulate a mixture of vMF distributions with given parameters. The algorithm in Figure 3 makes use of the algorithm in Figure 4.

```
**function** vsamp($\boldsymbol{\mu}, \kappa, n$)
{Adapted from [Woo94]}
In: $\boldsymbol{\mu}$ mean vector for vMF, $\kappa$ parameter for vMF
In: $n$, number of points to generate
Out: $S$ the Set of $n$ vMF($\boldsymbol{\mu}, \kappa$) samples

1. $d \leftarrow dim(\boldsymbol{\mu})$
2. $t_1 \leftarrow \sqrt{4\kappa^2 + (d-1)^2}$
3. $b \leftarrow (-2\kappa + t_1)/(d-1)$
4. $x_0 \leftarrow (1-b)/(1+b)$
5. $S \leftarrow zeros(n, d)$

6. $m \leftarrow (d-1)/2$
7. $c \leftarrow \kappa x_0 + (d-1)\log(1-x_0^2)$
8. **for** $i \leftarrow 1$ **to** $n$
       $t \leftarrow -1000$
       $u \leftarrow 1$
       **while** $(t < \log(u))$
           $z \leftarrow \beta(m, m)$      {$\beta(x, y)$ gives a beta random variable}
           $u \leftarrow rand$         {$rand$ gives a uniformly distributed random number.}
           $w \leftarrow \frac{(1-(1+b)z)}{(1-(1-b)z)}$
           $t \leftarrow \kappa w + (d-1)\log(1-x_0 w)$
       **end**
       $\mathbf{v} \leftarrow urand(d-1)$      {$urand(p)$ gives a $p$-dim vector from unif. distr. on sphere.}
       $\mathbf{v} \leftarrow \mathbf{v}/\|\mathbf{v}\|$
       $S(i, 1:d-1) \leftarrow \sqrt{1-w^2}\mathbf{v}^T$
       $S(i, d) = w$
9. **end**

{ We now have $n$ samples from vMF($[0\ 0\ \ldots\ 1]^T, \kappa$) }
10. Perform an orthogonal transformation on each sample in $S$
    The transformation has to satisfy $Q\boldsymbol{\mu} = [0\ 0\ \ldots\ 1]^T$
11. **return** $S$.
```

Figure 4: Algorithm to simulate a vMF

5.2 Experiment 1

In this section we discuss briefly some experiments carried out with the aim of verifying the accuracy of m.l.e. for parameters of a single vMF distribution. A consequence of the experiments is the verification of the vMF simulation algorithm given in Figure 4.

5.2.1 Experiment 1.1

This experiment deals with estimating parametes of a three-dimensional vMF distribution. The results are are summarized in Table1. The true mean and true concentration are denoted by $\boldsymbol{\mu}$ and κ respectively, n denotes the number of samples and $\hat{\boldsymbol{\mu}}, \hat{\kappa}$ denote the estimated parameters. These results clearly indicate that the m.l.e. for κ and $\boldsymbol{\mu}$ are quite accurate, and in the presence of large amounts of sample data m.l.e. approximate the true parameters quite well. Note that the calculations for κ were done using an approximation, but that does not lead to too much inaccuracy.

$\boldsymbol{\mu}$	κ	n	$\hat{\boldsymbol{\mu}}^T\boldsymbol{\mu}$	$\hat{\kappa}$
[.7071 .7071 0]′	4	100	.9994	4.1568
[.7071 .7071 0]′	10	1000	.9998	10.4561
[.1543 .6172 .7715]′	15	1000	1.000	15.2949

Table 1: MLE for single vMF with $p = 3$

5.2.2 Experiment 1.2

This experiment is in similar vein to experiment 1.1, except that we tried it for 20-dimensional simulated data. Table 2 summarizes the results. These experiments lend confidence to our belief in both the simulation and the MLE. Next we shall discuss MLE for simulated mixtures of vMFs.

$\boldsymbol{\mu}$	κ	n	$\hat{\boldsymbol{\mu}}^T\boldsymbol{\mu}$	$\hat{\kappa}$
Random vector	10	100	0.9739	10.2989
Random vector	10	1000	0.9983	10.2506

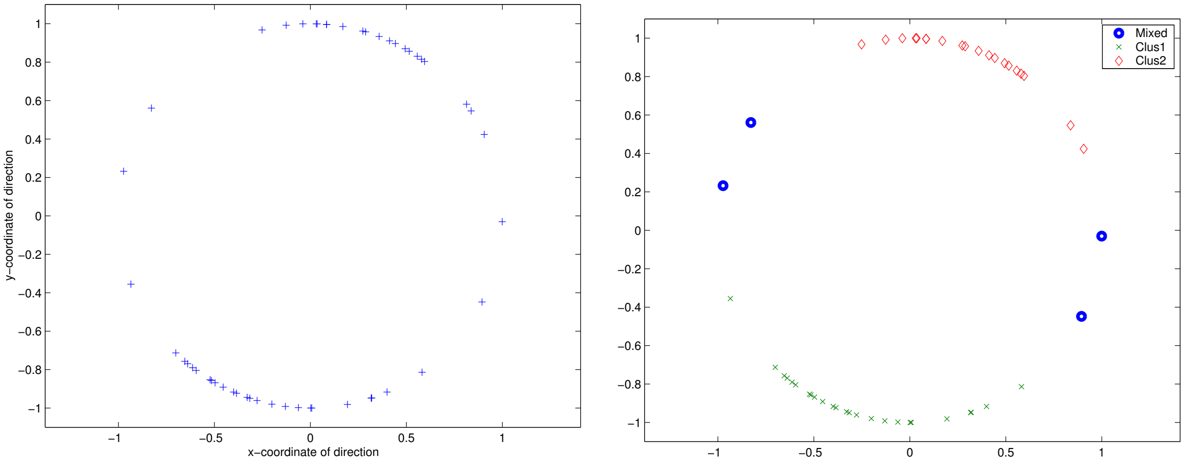
Table 2: MLE for vMF distribution with $p = 20$

5.3 Experiment 2

We provide a detailed example of clustering for a two component mixture of vMFs on a circle to illustrate the performance of our EM algorithm. The dataset that we considered was a small dataset of 50, two-dimensional points drawn from a mixture of two vMF distributions. The mean direction for each component was set to some random vector and κ was set to 4.

Figure 5(a) shows a plot of the points. From the plot we observe that there are two clusters of points (which is natural because the data was sampled from a mixture of two vMFs). Most points belong to either one component or the other. Some of the points seem to have mixed membership to each component. As we shall soon see, our EM algorithm figures out these points and assigns them fractionally to either component. The components as recovered by EM algorithm are illustrated in Figure 5(b).

From Figure 5(b), we can see that the points that we would have visually assigned to both components, have been given a mixed membership. This assignment seems to concur with our notion of a "correct" assignment. More precisely, in Figure 5(b), a point that has a probability exceeding 0.10, of membership to either component, is called a point with mixed membership. Thus



(a) Dataset for experiment 2.

(b) Components recovered by EM.

Figure 5: Illustration of experiment 2.

we have been able to recover both components and also glean useful information about points with a mixed membership.

5.4 Experiment 3

For this experiment we simulated three-dimensional directional data drawn from mixtures of two and three vMFs. The mean directions were chosen randomly as were the concentration parameters κ. For ease of presentation we have chosen to present results for 3 and higher dimensions in the tabular format given below. In Table 3, c denotes the number of components, $P(\omega)$ denotes the true

c	$P(\omega)$	$\hat{P}(\omega)$	avg $\hat{\boldsymbol{\mu}}^T\boldsymbol{\mu}$	avg $\frac{\lvert\hat{\kappa}-\kappa\rvert}{\kappa}$
2	0.55, 0.45	0.5424, 0.4576	0.9959	0.06
3	0.32 0.39 0.29	0.29 0.40 0.31	0.9965	0.05

Table 3: Parameter Estimation for vMF mixtures on a sphere.

priors and $\hat{P}(\omega)$ denotes estimated priors. Other symbols denote the usual quantities as described earlier. Though not included in this report we have available various other sets of simulated mixtures of vMFs on the sphere (for $c = 5, 6, 10, 12, 15, 20$) and the results are similar in quality to the ones reported in Table 3.

5.5 Experiment 4

This experiment focused on running the EM algorithm to estimate the parameters of a mixture of vMF distributions on a hypersphere ($p = 20$) where the number of components in the mixture was 5. From Table 4, it seems that the estimated priors and means are quite good. The estimated κ's do not seem to be that good. This behavior seems to be a manifestation of the approximation for kappa and a fallout of the limited number of simulated data points. Various problems arising due to numerical difficulties, compound the error as the algorithm progresses leading to somewhat unsatisfactory results.

Cluster	$P(\omega)$	$\hat{P}(\omega)$	κ	$\hat{\kappa}$	$\boldsymbol{\mu}^T \boldsymbol{\mu}$
1	0.165	0.193	11.1	14.1	0.89
2	0.200	0.192	8.6	9.1	0.85
3	0.185	0.181	7.0	9.9	0.87
4	0.210	0.204	8	7.3	0.85
5	0.240	0.230	15.0	15.5	0.96

Table 4: Estimated parameters for mixture of 5 vMFs with $p = 20$

5.6 Experiment 5

This experiment focused on running the EM algorithm to learn the parameters of a mixture of vMF distributions on a hypersphere ($p = 20$) where the number of components in the mixture was 20. The results of estimated parameters for Experiment 5 as given in Table 5 are reasonably good.

$\max \boldsymbol{\mu}^T \hat{\boldsymbol{\mu}}$	$\text{avg}\, \boldsymbol{\mu}^T \hat{\boldsymbol{\mu}}$	$\max \frac{\lvert\kappa-\hat{\kappa}\rvert}{\lvert\kappa\rvert}$	$\text{avg} \frac{\lvert\kappa-\hat{\kappa}\rvert}{\lvert\kappa\rvert}$	$\max \frac{\lvert P(\omega)-\hat{P}(\omega)\rvert}{\lvert P(\omega)\rvert}$	$\text{avg} \frac{\lvert P(\omega)-\hat{P}(\omega)\rvert}{\lvert P(\omega)\rvert}$
0.978	0.913	0.043	0.037	0.053	0.036

Table 5: Estimated parameters for Experiment 5

5.7 Experiment 6

We simulated a mixture with 5000 points, each in 1000 dimensions and having 4 components. The mean direction of each component was set to some random vector, and κ for each component was also set to a random number of in the range $[p/2..2p]$ ($p = 1000$). The mixing weights for each component were: $(.2576, .2440, .2398, .2586)$.

$\max \boldsymbol{\mu}^T \hat{\boldsymbol{\mu}}$	$\text{avg}\, \boldsymbol{\mu}^T \hat{\boldsymbol{\mu}}$	$\max \frac{\lvert\kappa-\hat{\kappa}\rvert}{\lvert\kappa\rvert}$	$\text{avg} \frac{\lvert\kappa-\hat{\kappa}\rvert}{\lvert\kappa\rvert}$	$\max \frac{\lvert P(\omega)-\hat{P}(\omega)\rvert}{\lvert P(\omega)\rvert}$	$\text{avg} \frac{\lvert P(\omega)-\hat{P}(\omega)\rvert}{\lvert P(\omega)\rvert}$
0.999	0.998	0.003	0.002	0.002	0.001

Table 6: Performance for Experiment 6

These results seem to be excellent and we do not notice tremendous errors for this seemingly complex dataset and the reason is that these results were produced by a ‘C’ implementation of our EM algorithm as described previously. The ‘C’ implementation uses extended precision arithmetic to overcome some of the numerical difficulties posed by the vMF distribution in high dimensions.

6 Conclusions

In this report we discussed the need for a directional model for certain types of data and proposed a mixture model capable of providing the needed model. The mixture model was a mixture of vMF distributions. We also saw how to compute the m.l.e. parameters for a single vMF distribution and mentioned the interesting numerical problems that arise when calculating kappa, the concentration parameter.

We derived m.l.e. equations for a mixture of vMF distributions and gave an EM algorithm to estimate m.l.e. MLE parameters. We verified our algorithms by running them on simulated data and described the technical difficulties encountered while doing so.

We observe that for data sampled from a mixture of von Mises-Fisher distributions we get fairly good estimates upon running our EM algorithm. For high dimensional data, numerical difficulties prevent us from getting very accurate results with a limited precision implementation. The extended precision implementation was able to get around these difficulties and yield excellent results even for high dimensional data.

One of the reasons is numerical difficulty in calculations due to Bessel functions of high order. The second, though more intrinsic difficulty is with the model itself. In very high dimensions we encounter very large values of κ. This leads to clustering decisions being made totally based on κ, something that is not desirable. Kappa captures the concentration, but we would prefer to give more preference to decisions based on the mean direction. Traditional spherical K-means already does that (though it must be noted again that it is a degenerate case of the more general vMF model). Also we note in passing that the "curse of dimensionality" leads to computational difficulties even in our case.

Further capabilities of the model that we have presented in this report need to be evaluated by applying to common domains like text clustering and gene expression data clustering. The investigation of such applications is part of our future work.

References

[AS74] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions.* Dover Publ. Inc., New York, 1974.

[Bil97] J. A. Bilmes. A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report ICSI-TR-97-021, University of Berkeley, 1997.

[DFG01] I. S. Dhillon, J. Fan, and Y. Guan. Efficient clustering of very large document collections. In V. Kumar R. Grossman, C. Kamath and R. Namburu, editors, *Data Mining for Scientific and Engineering Applications.* Kluwer Academic Publishers, 2001.

[DHS00] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley & Sons, 2000.

[DM01] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143–175, 2001.

[DMR03] I. S. Dhillon, E. M. Marcotte, and U. Roshan. Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, 2003. To appear.

[Hil81] G. W. Hill. Evaluation and inversion of the ratios of modified bessel functions. *ACM Transactions on Mathematical Software*, 7(2):199–208, June 1981.

[Knu98] D. E. Knuth. *The Art of Computer Programming*, volume 1: Fundamental Algorithms. Addison-Wesley, 3rd edition, 1998.

[MJ00] K. V. Mardia and P. Jupp. *Directional Statistics.* John Wiley and Sons Ltd., 2nd edition, 2000.

[vM18] R. von Mises. Über die "Ganzzahligkeit" der Atomgewichte und verwandte Fragen. *Phys. Z.*, 19:490–500, 1918.

[Wat96] G. N. Watson. *A treatise on the theory of Bessel functions.* Cambridge Mathematical Library. Cambridge University Press, 2nd (1944) edition, 1996.

[Woo94] A. T. A. Wood. Simulation of the von-Mises Distribution. *Communications of Statistics, Simulation and Computation*, 23:157–164, 1994.

A Mathematical background

To get a sound understanding of all the derivations performed in this report one needs some background. This Appendix provides the mathematical background required to derive some fundamental properties of the von Mises-Fisher distribution and also to supplement the understanding of the m.l.e. calculations for a mixture of vMFs.

A.1 Transformation to polar coordinates

Suppose we have an n vector $\mathbf{x}$ that we wish to translate to polar coordinates. We want to effect the transformation $\mathbf{x} = u(r, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{n-1})$ and $r = \|\mathbf{x}\|$. The co-ordinate transformation is generalized as below

$$\begin{aligned} x_k &\leftarrow r \sin\theta_1 \cdots \sin\theta_{k-1} \cos\theta_k, \qquad 1 \le k < n, \\ x_n &\leftarrow r \sin\theta_1 \cdots \sin\theta_{n-1}. \end{aligned}$$

We can re-express this generalization in the following inductive way: If $z_1, \ldots, z_{n-1}$ are the co-ordinates in $n-1$ dimensional space and $x_1, \ldots, x_n$ are the co-ordinates in n-dimensional space, and $\mathbf{x} = u(\boldsymbol{\theta})$ is the transformation to polar co-ordinates in $n-1$-space. Then we define (for $n > 3$):

$$\begin{aligned} x_i &= z_i \ \text{ for } \ 1 \le i \le n-2, \\ x_{n-1} &= z_{n-1} \cos\theta_{n-1}, \\ x_n &= z_{n-1} \sin\theta_{n-1}. \end{aligned}$$

The base case for the induction is $z_1 = r\cos\theta_1$ and $z_2 = r\sin\theta_1$. It is easy to verify that $\|\mathbf{x}\| = r$ and $\|\mathbf{z}\| = r$ as desired.

We know from vector calculus that: $d\mathbf{x} = |\det \mathbf{J}| d\boldsymbol{\theta}$ where $\mathbf{J}$ is the Jacobian matrix for the co-ordinate transformation and is given by

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta_1} & \cdots & \frac{\partial x_1}{\partial \theta_{n-1}} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta_1} & \cdots & \frac{\partial x_2}{\partial \theta_{n-1}} \\ \frac{\partial x_3}{\partial r} & \vdots & \vdots & \frac{\partial x_3}{\partial \theta_{n-1}} \\ \vdots & \cdots & \cdots & \vdots \\ \frac{\partial x_n}{\partial r} & \frac{\partial x_n}{\partial \theta_1} & \cdots & \frac{\partial x_n}{\partial \theta_{n-1}} \end{bmatrix}.$$

The determinant of the Jacobian of the transformation is (where $s_i = \sin\theta_i$ and $c_i = \cos\theta_i$):

$$|\mathbf{J}| = r^{n-1} \begin{vmatrix} c_1 & -s_1 & 0 & \cdots & 0 \\ s_1 c_2 & c_1 c_2 & -s_1 s_2 & \cdots & 0 \\ s_1 s_2 c_3 & c_1 s_2 c_3 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \prod_{i=1}^{n-1} s_i & c_1 \prod_{i=2}^{n-1} s_i & \cdots & \cdots & c_{n-1} \prod_{i=1}^{n-2} s_i \end{vmatrix}. \tag{A.1}$$

To calculate this determinant let us first define the following:

$$D_n(k) = \begin{vmatrix} c_k & -s_k & 0 & \cdots & 0 \\ s_k c_{k+1} & c_k c_{k+1} & -s_k s_{k+1} & \cdots & 0 \\ s_k s_{k+1} c_{k+2} & c_k s_{k+1} c_{k+2} & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \prod_{i=k}^{n+k-2} s_i & c_k \prod_{i=k+1}^{n+k-2} s_i & \cdots & c_t \prod_{\substack{i=k \\ i \ne t}}^{n+k-2} s_i & c_{n+k-2} \prod_{i=k}^{n+k-3} s_i \end{vmatrix}.$$

Then it is clear that the determinant of the Jacobian (A.1) is given by $r^{n-1}D_n(1)$. Expanding $D_n(k)$ along c_k and $-s_k$ and we get the following:

$$\begin{aligned} D_n(k) = c_k & \begin{vmatrix} c_k c_{k+1} & -s_k s_{k+1} & \cdots & 0 \\ c_k s_{k+1} c_{k+2} & -s_k s_{k+1} s_{k+2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ c_k \prod_{i=k+1}^{n+k-2} s_i & \cdots & \cdots & c_{n+k-2} \prod_{i=k}^{n+k-3} s_i \end{vmatrix} \\ + \quad s_k & \begin{vmatrix} s_k c_{k+1} & -s_k s_{k+1} & \cdots & 0 \\ s_k s_{k+1} c_{k+2} & -s_k s_{k+1} s_{k+2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \prod_{i=k}^{n+k-2} s_i & \cdots & \cdots & c_{n+k-2} \prod_{i=k}^{n+k-3} s_i \end{vmatrix}. \end{aligned} \tag{A.2}$$

Taking out c_k common from the first term and s_k from the second term and noting the fact that $c_k^2 + s_k^2 = 1$ we see that (A.2) reduces to:

$$D_n(k) = \begin{vmatrix} c_{k+1} & -s_k s_{k+1} & \cdots & 0 \\ s_{k+1} c_{k+2} & -s_k s_{k+1} s_{k+2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \prod_{i=k+1}^{n+k-2} s_i & \cdots & \cdots & c_{n+k-2} \prod_{i=k}^{n+k-3} s_i \end{vmatrix}.$$

Since all but the first column contain an s_k, we factor it out yielding

$$D_n(k) = s_k^{n-2} D_{n-1}(k+1). \tag{A.3}$$

Also by direct evaluation we know that $D_3(j) = s_j$. Hence on iterating (A.3) we conclude that:

$$D_n(k) = s_k^{n-2} s_{k+1}^{n-3} \cdots s_{k+n-3} = \prod_{j=k+1}^{n+k-2} s_{j-1}^{n+k-1-j}.$$

Since we know that $|J| = r^{n-1} D_n(1)$ we get the following:

$$|J| = r^{n-1} \prod_{j=2}^{n} \sin^{n-j} \theta_{j-1}.$$

For our case we have $r = 1$ and hence we can write:

$$dx_1\, dx_2 \,\cdots\, dx_n = \prod_{j=2}^{n} \sin^{n-j} \theta_{j-1} d\theta_{j-1}.$$

A.2 Some integrals and functions

In this section we gloss over some functions and integrals that prove to be useful while studying directional distributions.

A.2.1 The Gamma Function

We state and prove a few properties about the Gamma function that are useful for understanding some of the derivations associated with vMF distributions.

Leonhard Euler was the first to obtain the following generalization of the factorial function (see [Knu98])

$$n! = \lim_{m \to \infty} \frac{m^n m!}{(n+1)(n+2)\cdots(n+m)}.$$

A. M. Legendre introduced the notation: $n! = \Gamma(n+1) = n\Gamma(n)$ where

$$\Gamma(x) = \frac{x!}{x} = \lim_{m\to\infty} \frac{m^x m!}{x(x+1)(x+2)\cdots(x+m)}. \tag{A.4}$$

We now prove that

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt.$$

Our proof is based upon Ex. 1.2.5-19 of [Knu98]. We denote by $\Gamma_m(x)$ the quantity after the limit sign in equation (A.4) and we demonstrate that

$$\int_0^m \left(1 - t/m\right)^m t^{x-1}\, dt = \Gamma_m(x).$$

We can rewrite the integral above as (make the substitution $y = t/m$):

$$I_m(x) = \int_0^1 m^x (1-y)^m y^{x-1}\, dy.$$

Integrating by parts we find

$$\begin{aligned} I_m(x) &= m^x \left[\left. \frac{(1-y)^m y^x}{x} \right|_0^1 + \frac{m}{x} \int_0^1 (1-y)^{m-1} y^x\, dy \right], \\ &= \frac{m^{x+1}}{x} \int_0^1 (1-y)^{m-1} y^x\, dy. \end{aligned}$$

We can see that if we inductively assume

$$\Gamma_{m-1}(x) = m^x \int_0^1 (1-y)^{m-1} y^{x-1}\, dy,$$

then we may write $xI_m(x) = \Gamma_{m-1}(x+1)$. Hence using induction we can show that $I_m(x) = \Gamma_m(x)$. Now we note the fact that as $m \to \infty$, $(1 - t/m)^m \to e^{-t}$. This limit enables us to write

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt.$$

Note: The proof that $x\Gamma_m(x) = \Gamma_{m-1}(x+1)$ follows easily from the defintion of $\Gamma_m(x)$.

From either this integral or from the limiting definition we can verify the familiar property: $\Gamma(x+1) = x\Gamma(x)$. An important special case that comes up quite often is the value of $\Gamma(1/2)$. We shall evaluate it directly here. First we need a lemma:

Lemma A.1

$$\int_0^\infty e^{-u^2}\, du = \sqrt{\pi}/2.$$

Proof The integral in question is,

$$\begin{aligned} I &= \int_{-\infty}^\infty e^{-u^2}\, du, \\ I^2 &= \int_{-\infty}^\infty \int_{-\infty}^\infty e^{-(u^2+v^2)}\, du\, dv. \end{aligned}$$

Now put $u = r\cos\theta$ and $v = r\sin\theta$. So we get,

$$\begin{aligned} I^2 &= \int_0^{2\pi} d\theta \int_0^{\infty} e^{-r^2} r\,dr, \\ &= \pi. \end{aligned}$$

Hence, $I = \sqrt{\pi}$. It is easy to see that the integral under consideration is just half of I hence the lemma is true. ▮

Now we calculate the value of $\Gamma(1/2)$.

$$\Gamma(1/2) = \int_0^{\infty} e^{-t} t^{-1/2}\,dt.$$

We substitute $t = u^2$ so that $dt = 2u\,du$. Hence we have:

$$\Gamma(1/2) = \int_0^{\infty} e^{-u^2}\frac{1}{u}2udu = 2\int_0^{\infty} e^{-u^2}\,du.$$

But using Lemma A.1 we can conclude that: $\Gamma(1/2) = \sqrt{\pi}$.

A.2.2 The $\sin^n x$ integral

There are two ways that we show how to evaluate the following integral:

$$\int_0^{\pi} \sin^n x\,dx, \quad n > -1.$$

The first way is to type: `Integrate[Sin[x]^n,{x,0,π}]` in Mathematica and it will give back the answer:

$$\frac{\sqrt{\pi}\,\Gamma(\frac{1+n}{2})}{\Gamma(1+\frac{n}{2})}.$$

If one does not have recourse to Mathematica or some other such symbolic algebra system we could perform the integration as shown below.

$$\begin{aligned} I_n &= \int_0^{\pi} \sin^n x\,dx \\ &= \left[-\sin^{n-1} x\cos x\right]_0^{\pi} + (n-1)\int_0^{\pi} \sin^{n-2} x\cos^2 x\,dx \\ &= (n-1)\int_0^{\pi} \sin^{n-2} x\,dx - (n-1)\int_0^{\pi} \sin^n x\,dx \\ nI_n &= (n-1)I_{n-2}. \end{aligned}$$

By direct calculation we know that $I_2 = \pi/2$. So upon iteration we find out that

$$I_n = \frac{(n-1)(n-3)\cdots(n-2k+1)}{n(n-2)\cdots(n-2k)}\frac{\pi}{2} \quad ; \quad 2k = n-2.$$

We can write this as:

$$I_n = \frac{(\frac{n-1}{2})(\frac{n-3}{2})\cdots(\frac{n-2k+1}{2})}{\frac{n}{2}(\frac{n-2}{2})\cdots(\frac{n-2k}{2})}\frac{\pi}{2} \quad ; \quad 2k = n-2.$$

Since we know that $\Gamma(x+1) = x\Gamma(x)$ we can write

$$I_n = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n+2}{2})}\frac{\pi}{2\times\Gamma(3/2)}.$$

Using the fact that $\Gamma(3/2) = \sqrt{\pi}/2$ we conclude that

$$I_n = \frac{\sqrt{\pi}\,\Gamma(\frac{1+n}{2})}{\Gamma(1+\frac{n}{2})}.$$

I_n as given by the above equation is defined only for $n > -1$.

A.2.3 Useful formulae

The following differential equation gives rise to these modified Bessel functions:

$$z^2 w''(z) + zw'(z) - (z^2 + r^2)w(z) = 0. \tag{A.5}$$

This equation has solutions of the form: $w(z) = c_1 I_r(z) + c_z K_r(z)$ where $K_r(z)$ is the modified Bessel Function of the second kind.

The following two recurrence relations involving the derivative of the Bessel function are very useful in practice.

$$\kappa I'_p(\kappa) = pI_p(\kappa) + \kappa I_{p+1}(\kappa), \tag{A.6}$$

$$\kappa I'_p(\kappa) = \kappa I_{p-1}(\kappa) - pI_p(\kappa). \tag{A.7}$$

A standard definition of the modified Bessel function of order p and argument κ is

$$I_p(\kappa) = \sum_{r\geq 0} \frac{1}{\Gamma(p+r+1)r!}\left(\frac{\kappa}{2}\right)^{2r+p}. \tag{A.8}$$

Yet another definition is

$$I_p(\kappa) = \frac{2^{-p}\kappa^p}{\Gamma(p+1/2)\Gamma(1/2)} \int_0^\pi e^{\kappa\cos\theta} \sin^{2p}\theta d\theta, \tag{A.9}$$

which is equivalent to

$$I_p(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} \cos p\theta e^{\kappa\cos\theta} d\theta. \tag{A.10}$$

Finally, we can also write the above is a form that might be suitable for numerical integration procedures as follows:

$$I_p(\kappa) = \frac{2^{-p}\kappa^p}{\Gamma(p+1/2)\Gamma(1/2)} \int_{-1}^1 e^{\kappa t}(1-t^2)^{p-1/2} dt. \tag{A.11}$$

The following ratio is of principal importance to us,

$$A_p(\kappa) = \frac{I_{p/2}}{I_{p/2-1}}. \tag{A.12}$$

Another equivalent form can be derived using (A.11),

$$A_p(\kappa) = \frac{\int_{-1}^1 t^2 e^{\kappa t}(1-t^2)^{(p-3)/2} dt}{\int_{-1}^1 e^{\kappa t}(1-t^2)^{(p-3)/2} dt}. \tag{A.13}$$

We can obtain the following asymptotic representation for $A_p(\kappa)$,

$$\begin{gathered}\frac{\kappa}{p} - \frac{\kappa^3}{p^2\,(2+p)} + \frac{2\,\kappa^5}{p^3\,(2+p)\,(4+p)} + \frac{(-12-5\,p)\,\kappa^7}{p^4\,(2+p)^2\,(4+p)\,(6+p)} + \\ \frac{2\,(24+7\,p)\,\kappa^9}{p^5\,(2+p)^2\,(4+p)\,(6+p)\,(8+p)} + O(\kappa)^{10};\end{gathered} \tag{A.14}$$

in fact we can write $A_p(\kappa)$ as a convergent power series if $\kappa/p < 1$.

We are also interested in the derivative of $A_p(\kappa)$. We claim that the derivative of $A_p(\kappa)$, w.r.t. κ is given by,

$$A_p'(\kappa) = 1 - A_p(\kappa)^2 - \frac{p-1}{\kappa}A_p(\kappa). \tag{A.15}$$

Proof Let $s = p/2 - 1$. Then we have,

$$A_p'(\kappa) = \frac{I_{s+1}'(\kappa)}{I_s(\kappa)} - \frac{I_{s+1}(\kappa)}{I_s(\kappa)}\frac{I_s'(\kappa)}{I_s(\kappa)}. \tag{A.16}$$

Now we make use of (A.7) to obtain

$$\frac{I_{s+1}'(\kappa)}{I_s(\kappa)} = 1 - \frac{s+1}{\kappa}\frac{I_{s+1}(\kappa)}{I_s(\kappa)}, \tag{A.17}$$

and we use (A.6) to write

$$\frac{I_s'(\kappa)}{I_s(\kappa)} = \frac{s}{\kappa} + \frac{I_{s+1}(\kappa)}{I_s(\kappa)}. \tag{A.18}$$

We know that $A_p(\kappa) = \frac{I_{s+1}(\kappa)}{I_s(\kappa)}$ hence we conclude that

$$A_p'(\kappa) = 1 - A_p(\kappa)^2 - \left(\frac{s}{\kappa} + \frac{s+1}{\kappa}\right)A_p(\kappa). \tag{A.19}$$

Putting in $p/2 - 1$ for s we get the desired conclusion. ▮

If we invert the power series representation for $A_p(\kappa)$ we obtain the following approximation for κ:

$$p\,\bar{R}\left(1 + \frac{p\,\bar{R}^2}{2+p} + \frac{p^2\,(8+p)\,\bar{R}^4}{(2+p)^2\,(4+p)} + \frac{p^3\,(120+p\,(14+p))\,\bar{R}^6}{(2+p)^3\,(4+p)\,(6+p)} + \frac{p^4\,(24+p)\,(448+p\,(112+p\,(6+p)))\,\bar{R}^8}{(2+p)^4\,(4+p)^2\,(6+p)\,(8+p)}\right). \tag{A.20}$$

This estimate for κ does not really take into account the dimensionality of the data and thus for high dimensions (when κ is big by itself but κ/p is not very small or very big) it fails to yield accurate approximations. Note that $A_p(\kappa)$ is a ratio of Bessel functions that differ in their order by just one, so we can use a well known continued fraction expansion for representing $A_p(\kappa)$. For notational simplicity let us write the continued fraction expansion as:

$$A_{2s+2}(\kappa) = \frac{I_{s+1}}{I_s} = \frac{1}{\frac{2(s+1)}{\kappa}+}\,\frac{1}{\frac{2(s+2)}{\kappa}+}\cdots\,. \tag{A.21}$$

The continued fraction on the right is well known [Wat96]. Equation (A.21) and $A_p(\kappa) = \bar{R}$, allow us to write:

$$\frac{1}{\bar{R}} \approx \frac{2(s+1)}{\kappa} + \bar{R}.$$

Thus we can solve for κ to obtain the approximation,

$$\kappa \approx \frac{(2s+2)\bar{R}}{\bar{R} - \bar{R}^2}. \tag{A.22}$$

Since we made an approximation above, we incur some error, so we add a correction term (determined empirically) to the approximation of κ and obtain Equation (A.23),

$$\hat{\kappa} = \frac{\bar{R}p - \bar{R}^3}{1 - \bar{R}^2} \tag{A.23}$$

The above approximation can be generalized to include higher order terms in $\bar{R}$ to yield more accurate answers.[4] For $p = 2, 3$ highly accurate approximations can be found in [Hil81]. In most cases this estimate for κ is good enough because as far as inference is concerned a very accurate estimate does not buy us much. (The relative error between the true κ and this estimate has been found to be consistently lower than 0.05%)

For solving $A_p(\kappa) - \bar{R} = 0$, we can use (A.15) in any numerical method that may require the evaluation of the derivative of $A_p(\kappa)$ (such as Newton's method). In practice however, for very high dimensions (large p), Newton's iteration does not work that well because of numerical difficulties. In such cases, one could resort to other methods for improving the solution. Note that, again for practical purposes one does not really need very accurate calculations of κ. It is more of an academic numerical problem to find a good κ using an efficient and accurate root finder.

B Directional Distributions

The developments in this section are dependent upon the material presented in Appendix A. Thus, the reader who has not read Appendix A is advised to at least glance through it before proceeding with this appendix. Following the treatment in [MJ00], we will denote the probability element of $\mathbf{x}$ on a unit hyper-sphere by dS^{p-1}. The Jacobian of the transformation from (r, θ) to $\mathbf{x}$ is given by

$$dS^{p-1} = a_p(\theta)\, d\theta, \tag{B.1}$$

where we have (see Appendix A),

$$a_p(\theta) = \prod_{j=2}^{p-1} \sin^{p-j} \theta_{j-1}. \tag{B.2}$$

B.1 Uniform Distribution

If a direction $\mathbf{x}$ is uniformly distributed on S^{p-1} (unit hyper-sphere) then its probability element is $c_p\, dS^{p-1}$. The p.d.f. of θ is given by: $c_p a_p(\theta)$ (See Appendix A for a proof). Now we know that

$$\int c_p a_p(\theta) d\theta = 1, \tag{B.3}$$

hence using (B.2) we can write this as

$$c_p \int_0^{2\pi} d\theta_{p-1} \prod_{j=2}^{p-1} \int_0^{\pi} \sin^{p-j} \theta_{j-1} = 1. \tag{B.4}$$

Using the fact that (for $n > 0$, see Appendix A for a proof)

$$\int_0^{\pi} \sin^n x\, dx = \frac{\sqrt{\pi}\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n+2}{2})}, \tag{B.5}$$

we can easily solve equation (B.4) to give

$$c_p = \frac{\Gamma(p/2)}{2\pi^{p/2}}.$$

[4] Note that if one really wants more accurate approximations, it is better to use (A.23) as a starting point and then perform a couple of Newton-Raphson iterations, because it is easy to evaluate $A_p'(\kappa) = 1 - A_p(\kappa)^2 - \frac{p-1}{\kappa} A_p(\kappa)$.

B.2 The von Mises-Fisher distribution

A unit random vector $\mathbf{x}$ is said to have $p-$variate *von Mises-Fisher* distribution if its p.e. is:

$$c_p(\kappa)e^{\kappa\boldsymbol{\mu}^T\mathbf{x}}\,dS^{p-1}, \quad \mathbf{x}\in S^{p-1}\subseteq\mathbb{R}^p. \tag{B.6}$$

Where $\|\boldsymbol{\mu}\| = 1$ and $\kappa \geq 0$. We will derive the value of c_p the normalizing constant using the fact that:

$$\int_{\mathbf{x}\in S^{p-1}} c_p(\kappa)e^{\kappa\boldsymbol{\mu}'\mathbf{x}}\,d\mathbf{x} = 1. \tag{B.7}$$

To evaluate the integral above we make the transformation $\mathbf{y} = \mathbf{Q}\mathbf{x}$, where $y_1 = \boldsymbol{\mu}^T\mathbf{x}$ and $\mathbf{Q}$ is an orthogonal transformation. $\mathbf{x} = \mathbf{Q}^{-1}\mathbf{y}$ so $d\mathbf{x} = |\frac{\partial}{\partial\mathbf{y}}\mathbf{Q}^{-1}\mathbf{y}|d\mathbf{y}$. But since $\mathbf{Q}$ is an orthogonal transformation we have $d\mathbf{x} = d\mathbf{y}$. It is easy to see that the first row of the matrix $\mathbf{Q}$ is $\boldsymbol{\mu}^T$. We now make the transformation to polar co-ordinates: $\mathbf{y} = u(\boldsymbol{\theta})$. Using Equations (B.1) and (B.2) we can rewrite the integral above as:

$$\int_0^{2\pi} d\theta_{p-1}\int_0^{\pi} e^{\kappa\cos\theta_1}\sin^{p-2}\theta_1 d\theta_1 \prod_{j=3}^{p-1}\int_0^{\pi}\sin^{p-j}\theta_{j-1}d\theta_{j-1}. \tag{B.8}$$

Using Eq. (B.5) we can rewrite the above integral as:

$$I = 2\pi \times J_1 \times \pi^{\frac{p-3}{2}}\frac{\Gamma(\frac{p-2}{2})}{\Gamma(\frac{p-1}{2})}\frac{\Gamma(\frac{p-3}{2})}{\Gamma(\frac{p-2}{2})}\cdots\frac{\Gamma(1)}{\Gamma(\frac{3}{2})}, \tag{B.9}$$

where J_1 is given by:

$$J_1 = \int_0^{\pi} e^{\kappa\cos\theta_1}\sin^{p-2}\theta_1 d\theta_1. \tag{B.10}$$

But we know from (A.9) that:

$$I_{\frac{p-2}{2}}(\kappa) = \left(\frac{\kappa}{2}\right)^{\frac{p-2}{2}}\frac{J_1}{\Gamma(\frac{p-1}{2})\Gamma(\frac{1}{2})}. \tag{B.11}$$

Hence on combining (B.9) and (B.11) and using the fact that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ we see that the integral under question evaluates to:

$$I = \frac{(2\pi)^{p/2}}{\kappa^{p/2-1}}I_{p/2-1}(\kappa), \tag{B.12}$$

where $I_r(\kappa)$ is the modified Bessel Function as given by Eq. (A.9). We see that $c_p(\kappa) = I^{-1}$ and hence we have:

$$c_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2}I_{p/2-1}(\kappa)}. \tag{B.13}$$