Chapter 1

Directional Statistics in Machine Learning: a Brief Review

Abstract

The modern data analyst must cope with data encoded in various forms, vectors, matrices, strings, graphs, or more. Consequently, statistical and machine learning models tailored to different data encodings are important. We focus on data encoded as normalized vectors, so that their "direction" is more important than their magnitude. Specifically, we consider high-dimensional vectors that lie either on the surface of the unit hypersphere or on the real projective plane. For such data, we briefly review common mathematical models prevalent in machine learning, while also outlining some technical aspects, software, applications, and open mathematical challenges.

1.1 Introduction

Data are often represented as vectors in a Euclidean space $\mathbb{R}^p$, but frequently, data possess more structure and treating them as Euclidean vectors may be inappropriate. A simple example of this instance is when data are normalized to have unit norm, and thereby put on the surface of the *unit hypersphere* $\mathbb{S}^{p-1}$. Such data are better viewed as objects on a manifold, and when building mathematical models for such data it is often advantageous to exploit the geometry of the manifold (here $\mathbb{S}^{p-1}$).

For example, in classical information retrieval it has been convincingly demonstrated that cosine similarity is a more effective measure of similarity for analyzing and clustering text documents than just Euclidean distances. There is substantial empirical evidence that normalizing the data vectors helps to remove the biases induced by the length of a document and provide superior results [36, 37]. On a related note, the spherical k-means (`spkmeans`) algorithm [16] that runs k-means with cosine similarity for clustering unit norm vectors, has been found to work well for text clustering and a variety of other data. Another widely used similarity measure is *Pearson correlation*: given $x, y \in \mathbb{R}^d$ this defined as $\rho(x,y) := \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2}\times\sqrt{\sum_i (y_i-\bar{y})^2}}$, where $\bar{x} = \frac{1}{d}\sum_i x_i$ and $\bar{y} = \frac{1}{d}\sum_i y_i$. Mapping $x \mapsto \tilde{x}$ with $\tilde{x}_i = \frac{x_i-\bar{x}}{\sqrt{\sum_i (x_i-\bar{x})^2}}$ (similarly define $\tilde{y}$), we obtain the inner-product $\rho(x,y) = \langle \tilde{x}, \tilde{y} \rangle$. Moreover, $\|\tilde{x}\| = \|\tilde{y}\| = 1$. Thus, the Pearson correlation is exactly the cosine similarity between $\tilde{x}$ and $\tilde{y}$. More broadly, domains where similarity measures such as cosine, Jaccard or Dice [33] are more effective than measures derived from Mahalanobis type distances, possess intrinsic “directional” characteristics, and are hence better modeled as directional data [26].

This chapter recaps basic statistical models for *directional data*, which herein refers to unit norm vectors for which “direction” is more important than “magnitude.” In particular, we recall some basic distributions on the unit hypersphere, and then discuss two of the most commonly used ones: the von Mises-Fisher and Watson distributions. For these distributions, we describe maximum likelihood estimation as well as mixture modeling via the Expectation Maximization (EM) algorithm. In addition, we include a brief pointer to recent literature on applications of directional statistics within machine learning and related areas.

We warn the advanced reader that no new theory is developed in this chapter, and our aim herein is to merely provide an easy introduction. The material of this chapter is based on the author’s thesis [38], and the three papers [5, 39, 40], and the reader is referred to these works for a more detailed development and additional experiments.

1.2 Basic Directional Distributions

1.2.1 Uniform distribution

The probability element of the uniform distribution on $\mathbb{S}^{p-1}$ equals $c_p d\mathbb{S}^{p-1}$. The normalization constant c_p ensures that $\int_{\mathbb{S}^{p-1}} c_p d\mathbb{S}^{p-1} = 1$, from which it follows that

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

where $\Gamma(s) := \int_0^\infty t^{s-1}e^{-t}dt$ is the well-known Gamma function.

1.2.2 The von Mises-Fisher distribution

The vMF distribution is one of the simplest distributions for directional data and it has properties analogous to those of the multivariate Gaussian on $\mathbb{R}^p$. For instance, the maximum entropy density on $\mathbb{S}^{p-1}$ subject to the constraint that $E[x]$ is fixed, is a vMF density (see e.g., [32, pp. 172–174] and [27]).

A unit norm vector x has the von Mises-Fisher (vMF) distribution if its density is

$$p_{\text{vmf}}(x;\mu,\kappa) := c_p(\kappa)e^{\kappa\mu^T x},$$

where $\|\mu\| = 1$ and $\kappa \geq 0$. Integrating using polar coordinates, it can be shown [38, App. B.4.2] that the normalizing constant is given by

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

where $I_s(\kappa)$ denotes the modified Bessel function of the first kind [1].[1]

The vMF density $p_{\text{vmf}} = c_p(\kappa)e^{\kappa\mu^T x}$ is parameterized by the mean direction μ, and the *concentration* parameter κ, so-called because it characterizes how strongly the unit vectors drawn according to p_{vmf} are concentrated about the mean direction μ. Larger values of κ imply stronger concentration about the mean direction. In particular when $\kappa = 0$, p_{vmf} reduces to the uniform density on $\mathbb{S}^{p-1}$, and as $\kappa \to \infty$, p_{vmf} tends to a point density.

1.2.3 Watson distribution

The uniform and the vMF distributions are defined over *directions*. However, sometimes the observations are *axes* of direction, i.e., the vectors $\pm x \in \mathbb{S}^{p-1}$ are equivalent. This constraint is also denoted by $x \in \mathbb{P}^{p-1}$, where $\mathbb{P}^{p-1}$ is the projective hyperplane of dimension $p-1$. The multivariate Watson distribution [28] models such data; it is parametrized by a *mean-direction* $\mu \in \mathbb{P}^{p-1}$, and a *concentration* parameter $\kappa \in \mathbb{R}$, with probability density

$$p_{\text{wat}}(x;\mu,\kappa) := d_p(\kappa)e^{\kappa(\mu^T x)^2}, \qquad x \in \mathbb{P}^{p-1}. \tag{1.2.1}$$

The normalization constant $d_p(\kappa)$ is given by

$$d_p(\kappa) = \frac{\Gamma(p/2)}{2\pi^{p/2}M(\frac{1}{2},\frac{p}{2},\kappa)}, \tag{1.2.2}$$

where M is the confluent hypergeometric function [18, formula 6.1(1)]

$$M(a,c,\kappa) = \sum_{j\geq 0} \frac{a^{\overline{j}}}{c^{\overline{j}}}\frac{\kappa^j}{j!}, \qquad a,c,\kappa \in \mathbb{R}, \tag{1.2.3}$$

[1] Note that sometimes in directional statistics literature, the integration measure is normalized by the uniform measure, so that instead of $c_p(\kappa)$, one uses $c_p(\kappa)2\pi^{p/2}/\Gamma(p/2)$.

and $a^{\overline{0}} = 1$, $a^{\overline{j}} = a(a+1)\cdots(a+j-1)$, $j \geq 1$, denotes the *rising-factorial*.

Observe that for $\kappa > 0$, the density concentrates around μ as κ increases, whereas for $\kappa < 0$, it concentrates around the great circle orthogonal to μ.

1.2.4 Other distributions

We briefly summarize a few other interesting directional distributions, and refer the reader to [28] for a more thorough development.

Bingham distribution. Some axial data do not exhibit the rotational symmetry of Watson distributions. Such data could be potentially modeled using Bingham distributions, where the density at a point x is $B_p(x;K) := c_p(K)e^{x^T Kx}$, where the normalizing constant e_p can be shown to be $c_p(K) = \frac{\Gamma(p/2)}{2\pi^{p/2}M(\frac{1}{2},\frac{p}{2},K)}$, where $M(\cdot,\cdot,K)$ denotes the confluent hypergeometric function of matrix argument [31].

Note that since $x^T(K+\delta I_p)x = x^T Kx + \delta$, the Bingham density is identifiable only up to a constant diagonal shift. Thus, one can assume $\mathrm{Tr}(K) = 0$, or that the smallest eigenvalue of K is zero [28]. Intuitively, one can see that the eigenvalues of K determine the axes around which the data clusters, e.g., greatest clustering will be around the axis corresponding to the leading eigenvector of K.

Bingham-Mardia distribution. Certain problems require rotationally symmetric distributions that have a 'modal ridge' rather than just a mode at a single point. To model data with such characteristics [28] suggest a density of the form

$$p(x;\mu,\kappa,\nu) = c_p(\kappa)e^{\kappa(\mu^T x-\nu)^2}, \tag{1.2.4}$$

where as usual $c_p(\kappa)$ denotes the normalization constant.

Fisher-Watson distributions This distribution is a simpler version of the more general Fisher-Bingham distribution [28]. The density is

$$p(x;\mu,\mu_0,\kappa,\kappa_0) = c_p(\kappa_0,\kappa,\mu_0^T\mu)e^{\kappa_0\mu_0^T x+\kappa(\mu^T x)^2}. \tag{1.2.5}$$

Fisher-Bingham. This is a more general directional distribution; its density is

$$p(x;\mu,\kappa,A) = c_p(\kappa,A)e^{\kappa\mu^T x+x^T Ax}. \tag{1.2.6}$$

There does not seem to exist an easy integral representation of the normalizing constant, and in an actual application one needs to resort to some sort of approximation for it (such as a saddle-point approximation). Kent distributions arise by putting an additional constraint $A\mu = 0$ in (1.2.6).

1.3 Related work and applications

The classical references on directional statistics are [26, 27, 46]; a more recent, updated reference is [28]. Additionally, for readers interested in statistics on manifolds, a good starting point is [11]. To our knowledge, the first work focusing on high-dimensional application of directional statistics was [5], where the key application was clustering of text and gene expression data using mixtures of vMFs. There exist a vast number of other applications and settings where hyperspherical or manifold data arise. Summarizing all of these is clearly beyond the scope of this chapter. We mention below a smattering of some works that are directly related to this chapter.

We note a work on feature extraction based on correlation in [19]. Classical data mining applications such as topic modeling for normalized data are studied in [4,34]. A semi-parametric setting using Dirichlet process mixtures for spherical data is [41]. Several directional data clustering settings include: depth images using Watson mixtures [20]; a k-means++ [3] style procedure for mixture of vMFs [29]; clustering on orthogonal manifolds [10]; mixtures of Gaussian and vMFs [23]. Directional data has also been used in several biomedical (imaging) applications, for example [30], fMRI [24], white matter supervoxel segmentation [9], and brain imaging [35]. In signal processing there are applications to spatial fading using vMF mixtures [25] and speaker modeling [44]. Finally, beyond vMF and Watson, it is worthwhile to consider the Angular Gaussian distribution [45], which has been applied to model natural images for instance in [22].

1.4 Modeling directional data: maximum-likelihood estimation

In this section we briefly recap data models involving vMF and Watson distributions. In particular, we describe maximum-likelihood estimation for both distributions. As is well-known by now, for these distributions estimating the mean μ is simpler than estimating the concentration parameter κ.

1.4.1 Maximum-Likelihood estimation for vMF

Let $\mathscr{X} = \{x_1, \ldots, x_n\}$ be a set of points drawn from $p_{\text{vmf}}(x; \mu, \kappa)$. We wish to estimate μ and κ by solving the m.l.e. optimization problem

$$\max \ell(\mathscr{X}; \mu, \kappa) := \log c_p(\kappa) + \textstyle\sum_{i=1}^n \kappa \mu^T x_i, \quad \text{s.t. } \|\mu\| = 1,\ \kappa \geq 0. \tag{1.4.1}$$

Writing $\frac{\|\sum_i x_i\|}{n} = \bar{r}$, a brief calculation shows that the optimal solution satisfies

$$\mu = \tfrac{1}{n\bar{r}} \textstyle\sum_{i=1} x_i, \quad \kappa = A_p^{-1}(\bar{r}), \tag{1.4.2}$$

where the nonlinear map A_p is defined as

$$A_p(\kappa) = \frac{-c'_p(\kappa)}{c_p(\kappa)} = \frac{I_{p/2}(\kappa)}{I_{p/2-1}(\kappa)} = \bar{r}. \tag{1.4.3}$$

The challenge is to solve (1.4.3) for κ. For small values of p (e.g., $p = 2,3$) the simple estimates provided in [28] suffice. But for machine learning problems, where p is typically very large, these estimates do not suffice. In [5], the authors provided efficient numerical estimates for κ that were obtained by truncating the continued fraction representation of $A_p(\kappa)$ and solving the resulting equation. These estimates were then corrected to yield the approximation

$$\hat{\kappa} = \frac{\bar{r}(p - \bar{r}^2)}{1 - \bar{r}^2}, \tag{1.4.4}$$

which turns out to be remarkably accurate in practice.

Subsequently, [43] showed simple bounds for κ by exploiting inequalities about the Bessel ratio $A_p(\kappa)$—this ratio possesses several nice properties, and is very amenable to analytic treatment [2]. The work of [43] lends theoretical support to the empirically determined approximation (1.4.4), by essentially showing this approximation lies in the "correct" range. Tanabe et al. [43] also presented a fixed-point iteration based algorithm to compute an approximate solution κ.

The *critical* difference between this approximation and the next two is that it does not involve any Bessel functions (or their ratio). That is, not a single evaluation of $A_p(\kappa)$ is needed—an advantage that can be significant in high-dimensions where it can be computationally expensive to compute $A_p(\kappa)$. Naturally, one can try to compute $\log I_s(\kappa)$ $(s = p/2)$ to avoid overflows (or underflows as the case may be), though doing so introduces yet another approximation. Therefore, when running time and simplicity are of the essence, approximation (1.4.4) is preferable.

Approximation (1.4.4) can be made more exact by performing a few iterations of Newton's method. To save runtime, [39] recommends only two-iterations of Newton's method, which amounts to computing κ_0 using (1.4.4), followed by

$$\kappa_{s+1} = \kappa_s - \frac{A_p(\kappa_s) - \bar{R}}{1 - A_p(\kappa_s)^2 - \frac{(p-1)}{\kappa_s} A_p(\kappa_s)}, \quad s = 0, 1. \tag{1.4.5}$$

Approximation (1.4.5) was shown in [39] to be competitive in running time with the method of [43], and was seen to be overall more accurate. Approximating κ remains a topic of research interest, as can be seen from the recent works [12, 21].

1.4.2 Maximum-Likelihood estimation for Watson

Let $\mathscr{X} = \{x_1, \ldots, x_n\}$ be i.i.d. samples drawn from $p_{\text{wat}}(x; \mu, \kappa)$. We wish to estimate μ and κ by maximizing the log-likelihood

$$\ell(\mathscr{X}; \mu, \kappa) = n\big(\kappa \mu^\top S \mu - \ln M(1/2, p/2, \kappa) + \gamma\big), \tag{1.4.6}$$

subject to $\mu^T \mu = 1$, where $S = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$ is the sample *scatter matrix*, and γ is a constant. Considering the first-order optimality conditions of (1.4.6) leads to the following parameter estimates [28, Sec. 10.3.2]

$$\hat{\mu} = \pm s_1 \quad \text{if} \quad \hat{\kappa} > 0, \qquad \hat{\mu} = \pm s_p \quad \text{if} \quad \hat{\kappa} < 0, \tag{1.4.7}$$

where $s_1, s_2, \ldots, s_p$ are (normalized) eigenvectors of the scatter matrix S corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

To estimate the concentration parameter $\hat{\kappa}$ we must solve:[2]

$$g(\tfrac{1}{2}, \tfrac{p}{2}; \hat{\kappa}) := \frac{\frac{\partial}{\partial\kappa} M(\frac{1}{2}, \frac{p}{2}, \hat{\kappa})}{M(\frac{1}{2}, \frac{p}{2}, \hat{\kappa})} \;=\; \hat{\mu}^\top S \hat{\mu} \;:=\; r \qquad (0 \leq r \leq 1), \tag{1.4.8}$$

Notice that (1.4.7) and (1.4.8) are coupled—so we simply solve both $g(1/2, p/2; \hat{\kappa}) = \lambda_1$ and $g(1/2, p/2; \hat{\kappa}) = \lambda_p$, and pick the solution that yields a higher log-likelihood.

The hard part is to solve (1.4.8). One could use a root-finding method (e.g. Newton-Raphson), but similar to the vMF case, an out-of-the-box root-finding approach can be unduly slow or numerically hard as data dimensionality increases. The authors of [40] consider the following more general equation:

$$\begin{gathered} g(a, c; \kappa) := \frac{M'(a, c; \kappa)}{M(a, c; \kappa)} = r \\ c > a > 0, \quad 0 \leq r \leq 1, \end{gathered} \tag{1.4.9}$$

and derive for it high-quality closed form numerical approximations. These approximations improve upon two previous approaches, that of [7] and [38]. Bijral et al. [7] followed the continued-fraction approach of [5] to obtain the heuristic approximation

$$BBG(r) := \frac{cr - a}{r(1-r)} + \frac{r}{2c(1-r)}. \tag{1.4.10}$$

Other heuristic approximations were presented by the author in [38].

The following theorem of [40] provides rigorously justified approximations, most of which are typically more accurate than previous heuristics.

Theorem 1.4.1 ([40])**.** *Let the solution to $g(a, c; \kappa) = r$ be denoted by $\kappa(r)$. Consider the following three bounds:*

$$(\textit{lower bound}) \qquad L(r) = \frac{rc - a}{r(1-r)} \left(1 + \frac{1-r}{c-a}\right), \tag{1.4.11}$$

$$(\textit{bound}) \qquad B(r) = \frac{rc - a}{2r(1-r)} \left(1 + \sqrt{1 + \frac{4(c+1)r(1-r)}{a(c-a)}}\right), \tag{1.4.12}$$

$$(\textit{upper bound}) \qquad U(r) = \frac{rc - a}{r(1-r)} \left(1 + \frac{r}{a}\right). \tag{1.4.13}$$

Let $c > a > 0$, and $\kappa(r)$ be the solution (1.4.9)*. Then, we have*

1. for $a/c < r < 1$,

$$L(r) < \kappa(r) < B(r) < U(r), \tag{1.4.14}$$

2. for $0 < r < a/c$,

$$L(r) < B(r) < \kappa(r) < U(r). \tag{1.4.15}$$

3. and if $r = a/c$, then $\kappa(r) = L(a/c) = B(a/c) = U(a/c) = 0$.

All three bounds (L, B, and U) are also asymptotically precise at $r = 0$ and $r = 1$.

[2] We need $\lambda_1 > \lambda_2$ to ensure a unique m.l.e. for positive κ, and $\lambda_{p-1} > \lambda_p$, for negative κ

1.5 Mixture models

Many times a single vMF or Watson distribution is insufficient to model data. In these cases, a richer model (e.g., for clustering, or as a generative model, etc.) such as a mixture model may be more useful. We summarize in this section mixtures of vMF (movMF) and mixtures of Watson (moW) distributions. The former was originally applied to high-dimensional text and gene expression data in [5], and since then it has been used in a large number of applications (see also Section 1.3). The latter has been applied to genetic data [40], as well as in other data mining applications [7].

Let $p(x;\mu,\kappa)$ denote either a vMF density or a Watson density. We consider mixture models of K different vMF densities or K different Watson densities. Thus, a given unit norm observation vector x has the *mixture density*

$$f\big(x;\{\mu_j\}_{j=1}^K,\{\kappa_j\}_{j=1}^K\big) := \sum\nolimits_{j=1}^K \pi_j p(x;\mu_j,\kappa_j). \tag{1.5.1}$$

Suppose we observe the set $\mathscr{X} = \{x_1,\ldots,x_n \in \mathbb{P}^{p-1}\}$ of i.i.d. samples drawn from (1.5.1). Our aim is to infer the mixture parameters $(\pi_j,\mu_j,\kappa_j)_{j=1}^K$, where $\sum_j \pi_j = 1$, $\pi_j \geq 0$, $\|\mu_j\| = 1$, and $\kappa_j \geq 0$ for an movMF and $\kappa_j \in \mathbb{R}$ for a moW.

1.5.1 EM algorithm

A standard, practical approach to estimating the mixture parameters is via the Expectation Maximization (EM) algorithm [14] applied to maximize the mixture log-likelihood for $\mathscr{X}$. Specifically, we seek to maximize

$$\ell(\mathscr{X};\{\pi_j,\mu_j,\kappa_j\}_{j=1}^K) := \sum\nolimits_{i=1}^n \ln\Big(\sum\nolimits_{j=1}^K \pi_j p(x;\mu_k,\kappa_j)\Big). \tag{1.5.2}$$

To apply EM, first we use Jensen's inequality to compute the lower bound

$$\ell(\mathscr{X};\{\pi_j,\mu_j,\kappa_j\}_{j=1}^K) \geq \sum\nolimits_{ij} \beta_{ij} \ln\left(\pi_j p(x_i|\mu_j,\kappa_j)/\beta_{ij}\right). \tag{1.5.3}$$

Then, the E-Step sets β_{ij} to the *posterior* probability (for x_i given component j):

$$\beta_{ij} := \frac{\pi_j p(x_i|\mu_j,\kappa_j)}{\sum_l \pi_l p(x_i|\mu_l,\kappa_l)}. \tag{1.5.4}$$

With this choice of β_{ij}, the M-Step maximizes (1.5.3), which is essentially just a maximum-likelihood problem, to obtain parameter updates. In particular, we obtain *M-Step for movMF:*

$$\mu_j = \frac{r_j}{\|r_j\|}, \quad r_j = \sum\nolimits_{i=1}^n \beta_{ij} x_i, \tag{1.5.5}$$

$$\kappa_j = A_p^{-1}(\bar{r}_j), \quad \bar{r}_j = \frac{\|r_j\|}{\sum_{i=1}^n \beta_{ij}}. \tag{1.5.6}$$

M-Step for moW:

$$\mu_j = s_1^j \quad \text{if} \quad \kappa_j > 0, \qquad \mu_j = s_p^j \quad \text{if} \quad \kappa_j < 0, \tag{1.5.7}$$

$$\kappa_j = g^{-1}(1/2, p/2, r_j), \quad \text{where} \quad r_j = \mu_j^\top S^j \mu_j \tag{1.5.8}$$

```
Input: $x = \{x_1, \ldots, x_n :$ where each $\|x_i\| = 1\}, K$
Output: Parameter estimates $\pi_j$, $\mu_j$, and $\kappa_j$, for $1 \leq j \leq K$
Initialize $\pi_j, \mu_j, \kappa_j$ for $1 \leq j \leq K$
while not converged do
    {Perform the E-step of EM}
    foreach i and j do
        Compute $\beta_{ij}$ using (1.5.4) (or via (1.5.9))
    end
    {Perform the M-Step of EM}
    for $j = 1$ to $K$ do
        $\pi_j \leftarrow \frac{1}{n}\sum_{i=1}^{n} \beta_{ij}$
        For movMF: compute $\mu_j$ and $\kappa_j$ using (1.5.5) and (1.5.6)
        For moW: compute $\mu_j$ and $\kappa_j$ using (1.5.7) and (1.5.8)
    end
end
```

Algorithm 1: EM Algorithm for movMF and moW

where s_1^j denotes the top eigenvector corresponding to eigenvalue $\lambda_i(S_j)$ of the *weighted-scatter matrix*

$$S^j = \frac{1}{\sum_{i=1}^{n} \beta_{ij}} \sum_{i=1}^{n} \beta_{ij} x_i x_i^T .$$

For both movMF and moW, the component probabilities are as usual $\pi_j = \frac{1}{n}\sum_i \beta_{ij}$. Iterating between (1.5.4) and the M-Steps we obtain an EM algorithm. Pseudo-code for such a procedure is shown below as Algorithm 1.

Hard Assignments. To speed up EM, we can replace can E-Step (1.5.4) by the standard *hard-assignment* rule:

$$\beta_{ij} = \begin{cases} 1, & \text{if } j = \operatorname{argmax}_{j'} \ln \pi_l + \ln p(x_i|\mu_l, \kappa_l), \\ 0, & \text{otherwise.} \end{cases} \tag{1.5.9}$$

The corresponding M-Step also simplifies considerably. Such hard-assignments maximize a lower-bound on the incomplete log-likelihood and yield *partitional-clustering* algorithms.

Initialization. For movMF, typically an initialization using spherical kmeans (spkmeans) [16] can be used. The next section presents arguments that explain why this initialization is natural for movMF. Similarly, for moW, an initialization based on diametrical kmeans [15] can be used, though sometimes even an spkmeans initialization suffices [40].

1.5.2 Limiting versions

It is well-known that the famous k-means algorithm may be obtained as a limiting case of the EM algorithm applied to a mixture of Gaussians. Analogously, the spherical kmeans algorithm of [16] that clusters unit norm vectors and finds unit norm means (hence 'spherical') can be viewed as the limiting case of a movMF. Indeed, assume that the priors of all mixture components are equal. Furthermore, assume that all the mixture components have equal concentration parameters κ and let $\kappa \to \infty$. Under these assumptions, the E-Step (1.5.9) reduces to assigning a point x_i to the cluster nearest to it, which here is given by the cluster with whose centroid the given point has largest dot product. In other words, a point x_i is assigned to cluster $k = \operatorname{argmax}_j x_i^T \mu_j$ because $\beta_{ik} \to 1$ and $\beta_{ij} \to 0$ for $j \neq k$ in (1.5.9).

In a similar manner, the diametrical clustering algorithm of [15] also may be viewed as a limiting case of EM applied to a moW. Recall that the diametrical clustering algorithm groups together correlated and anti-correlated unit norm data vectors into the same cluster, i.e., it treats diametrically opposite points equivalently. Remarkably, it turns out that the diametrical clustering algorithm of [15] can be obtained as follows: Let $\kappa_j \to \infty$, so that for each i, the corresponding posterior probabilities $\beta_{ij} \to \{0,1\}$; the particular β_{ij} that tends to 1 is the one for which $(\mu_j^\top x_i)^2$ is maximized in the E-Step; subsequently the M-Step (1.5.7), (1.5.8) also simplifies, and yields the same updates as made by the diametrical clustering algorithm.

Alternatively, we can obtain diametrical clustering from the hard-assignment heuristic of EM applied to a moW where all mixture components have the same (positive) concentration parameter κ. Then, in the E-Step (1.5.9), we can ignore κ altogether, which reduces Alg. 1 to the diametrical clustering procedure.

1.5.3 Application: clustering using movMF

Mixtures of vMFs have been successfully used in text clustering; see [6] for a detailed overview. We present below results of a two main experiments below: (i) simulated data; and (ii) Slashdot news articles.

The key characteristic of text data is its high dimensionality. And for modeling clusters of such data using a movMF, the approximate computation of the concentration parameter κ as discussed in Sec. 1.4.1 is of great importance: without this approximation, the computation breaks down due to floating point difficulties.

For (i), we simulate a mixture of 4 vMFs in with $p = 1000$ each, and draw a sample of 5000 data points. The clusters are chosen to be roughly of the same size and their relative mixing proportions are $(0.25, 0.24, 0.25, 0.26)$, with concentration parameters (to one digiit) $(651.0, 267.8, 267.8, 612.9)$, and random units vectors as means. This is the same data as the `big-mix` data in [6]. We generated the samples using the `vmfsamp` code (available from the author upon request).

For (ii), we follow [6] and use news articles from the Slashdot website. These articles are tagged and cleaned to retain 1000 articles that more clearly belong to a primary category / cluster. We report results on 'Slash-7' and 'Slash-6'; the first contains 6714 articles in 7 primary categories: business, education, entertainment, games, music, science, and internet; while the second contains 5182 articles in 6 categories: biotech, Microsoft, privacy, Google, security, and space.

Performance evaluation. There are several ways to evaluate performance of a clustering method. For the simulated data, we actually know the true parameters from which the dataset was simulated, hence we compare the error in estimated parameter values. For the Slashdot data, we have knowledge of "ground truth" labels, so we can use the *normalized mutual information (NMI)* [42], a measure that was also previously used to assess movMF based clustering [5,6]. Suppose the predicted cluster labels are $\hat{Y}$ and the true labels are Y, then the NMI between Y and $\hat{Y}$ is defined as

$$\mathrm{NMI}(Y,\hat{Y}) := \frac{I(Y,\hat{Y})}{\sqrt{H(Y)H(\hat{Y})}}, \tag{1.5.10}$$

where $I(\cdot,\cdot)$ denotes the usual mutual information and H denotes the entropy [13]. When the predicted labels agree perfectly with the true labels, then NMI equals 1; thus higher values of NMI are better.

Results on 'bigsim'. The results of the first experiment are drawn from [5], and are reported in Table 1.1. From the results it is clear that on this particular simulated data, EM manages to recover the true parameters to quite a high degree of accuracy. Part of this reason is due to the high values of the concentration parameter: as κ increases, the probability mass concentrates, which makes it easier to separate the clusters using EM. To compute the values in the table, we ran EM with soft-assignments and then after convergence used assignment (1.5.9).

$\min \mu^T\hat{\mu}$	$\max \frac{\lvert\kappa-\hat{\kappa}\rvert}{\kappa}$	$\max \frac{\lvert\pi-\hat{\pi}\rvert}{\pi}$
0.994	0.006	0.002

Table 1.1 *Accuracy of parameter estimation via EM for movMF on the 'bigsim' data. We report the worst values (the averages were better) seen across 20 different runs. For the estimated mean, we report the worst inner product with the true mean; for the concentration and mixture priors we report worst case relative errors.*

Results on Slashdot. Our experiment here reports performance of our implementation of Alg. 1 (EM for movMF) against Latent Dirichlet Allocation (LDA) [8] and a Exponential-family Dirichlet compound multinomial model (EDCM) [17]. For experiments (on other data) comparing movMF based clustering to spkmeans, we refer the reader to the extensive results in [5], as well as [6].

Table 1.2 reports results of comparing Alg. 1 specialized for movMFs against LDA and EDCM. As can be seen from the results, the vMF mixture leads to much higher quality clustering than the other two competing approaches. We did not test an optimized implementation (and used our own MATLAB implementation), but note anecdotally that the EM procedure was 3–5 times faster than the others.

Dataset	moVMF	LDA	ECDM
Slash-7	0.39	0.22	0.31
Slash-6	0.65	0.36	0.46

Table 1.2 *Comparison of NMI values of moVMF versus LDA and ECDM (derived from [6]).*

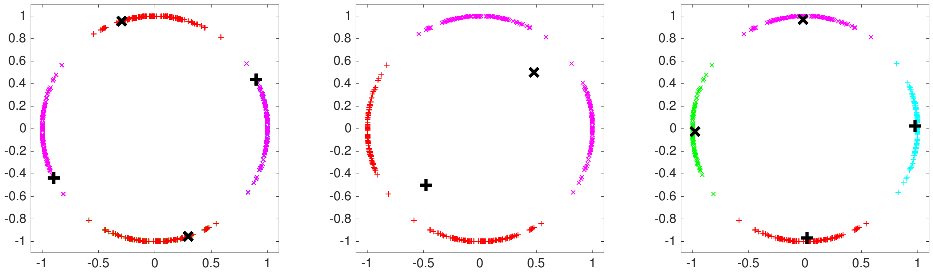

Figure 1.1 *The left panel shows axially symmetric data that has two clusters (centroids are indicated by '+' and 'x'). The middle and right panel shows clustering yielded by (Euclidean) K-means (note that the centroids fail to lie on the circle in this case) with $K = 2$ and $K = 4$, respectively. Diametrical clustering recovers the true clusters in the left panel.*

1.5.4 Application: clustering using moW

Figure 1.1 shows a toy example of axial data. Here, the original data has two clusters (leftmost panel of Fig. 1.1). If we cluster this data into two clusters using Euclidean kmeans, we obtain the plot in the middle panel; clustering into 4 groups using Euclidean kmeans yields the rightmost panel. As is clear from the figure, Euclidean kmeans cannot discover the desired structure, if the true clusters are on axial data. The diametrical clustering algorithm of [15] discovers the two clusters (leftmost panel), which also shows the mean vectors $\pm\mu$ for each cluster. Recall that as mentioned above, the diametrical clustering method is obtained as the limiting case of EM on moW.

1.6 Conclusion

We summarized a few distributions from directional statistics that are useful for modeling normalized data. We focused in particular on the von Mises-Fisher distribution (the "Gaussian" of the hypersphere) and the Watson distribution (for axially symmetric data). For both of these distributions, we recapped maximum likelihood parameter estimation as well as mixture modeling using the EM algorithm. For extensive numerical results on clustering using mixtures of vMFs, we refer the reader to the original paper [5]; similarly, for mixtures of Watsons please see [40]. The latter paper also describes asymptotic estimates of the concentration parameter in detail.

Now directional distributions are widely used in machine learning (Sec. 1.3 provides some pointers to related work), and we hope the brief summary provided in this chapter helps promote wider understanding about these. In particular, we hope to see more exploration of directional models in the following important subareas: Bayesian models, Hidden Markov Models using directional models, and deep generative models.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables*. Dover, New York, June 1974.

[2] D. E. Amos. Computation of modified Bessel functions and their ratios. *Mathematics of Computation*, 28(125):235–251, 1974.

[3] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[4] A. Banerjee and S. Basu. Topic models over text streams: A study of batch and online unsupervised learning. In *SDM*, volume 7, pages 437–442. SIAM, 2007.

[5] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *JMLR*, 6:1345–1382, Sep 2005.

[6] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Text Clustering with Mixture of von Mises-Fisher Distributions. In A. N. Srivastava and M. Sahami, editors, *Text Mining: Theory, Applications, and Visualization*. CRC Press, 2009.

[7] A. Bijral, M. Breitenbach, and G. Z. Grudic. Mixture of Watson Distributions: A Generative Model for Hyperspherical Embeddings. In *Artificial Intelligence and Statistics (AISTATS 2007)*, J. Machine Learning Research - Proceedings Track 2, pages 35–42, 2007.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[9] R. P. Cabeen and D. H. Laidlaw. White matter supervoxel segmentation by axial dp-means clustering. In *Medical Computer Vision. Large Data in Medical Imaging*, pages 95–104. Springer, 2014.

[10] H. Cetingul and R. Vidal. Intrinsic mean shift for clustering on stiefel and grassmann manifolds. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1896–1902. IEEE, 2009.

[11] Y. Chikuse. *Statistics on special manifolds*, volume 174. Springer Science & Business Media, 2012.

[12] D. Christie. Efficient von Mises-Fisher concentration parameter estimation using Taylor series. *Journal of Statistical Computation and Simulation*, 85(16):1–7, 2015.

[13] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons,

New York, USA, 1991.

[14] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society*, 39, 1977.

[15] I. S. Dhillon, E. M. Marcotte, and U. Roshan. Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, 19(13):1612–1619, 2003.

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

[17] C. Elkan. Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In *Proceedings of the 23rd international conference on Machine learning*, pages 289–296. ACM, 2006.

[18] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi. *Higher transcendental functions*, volume 1. McGraw Hill, 1953.

[19] Y. Fu, S. Yan, and T. S. Huang. Correlation metric for generalized feature extraction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2229–2235, 2008.

[20] M. A. Hasnat, O. Alata, and A. Trémeau. Unsupervised clustering of depth images using watson mixture model. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 214–219. IEEE, 2014.

[21] K. Hornik and B. Grün. On maximum likelihood estimation of the concentration parameter of von mises–fisher distributions. *Computational statistics*, 29(5):945–957, 2014.

[22] R. Hosseini. *Natural Image Modelling using Mixture Models with compression as an application*. PhD thesis, Berlin, Technische Universtität Berlin, Diss., 2012, 2012.

[23] P. Kasarapu and L. Allison. Minimum message length estimation of mixtures of multivariate Gaussian and von Mises-Fisher distributions. *Machine Learning*, pages 1–46, 2015.

[24] D. Lashkari, E. Vul, N. Kanwisher, and P. Golland. Discovering structure in the space of fmri selectivity profiles. *Neuroimage*, 50(3):1085–1098, 2010.

[25] K. Mammasis, R. W. Stewart, and J. S. Thompson. Spatial fading correlation model using mixtures of von mises fisher distributions. *Wireless Communications, IEEE Transactions on*, 8(4):2046–2055, 2009.

[26] K. V. Mardia. *Statistical Distributions in Scientific Work*, volume 3, chapter Characteristics of directional distributions, pages 365–385. Reidel, Dordrecht, 1975.

[27] K. V. Mardia. Statistics of directional data. *J. Royal Statistical Society. Series B (Methodological)*, 37(3):349–393, 1975.

[28] K. V. Mardia and P. Jupp. *Directional Statistics*. John Wiley and Sons Ltd., second edition, 2000.

[29] M. Mash'al and R. Hosseini. K-means++ for Mixtures of von Mises-Fisher Distributions. In *Information and Knowledge Technology (IKT), 2015 7th Con-*

ference on, pages 1–6. IEEE, 2015.

[30] T. McGraw, B. C. Vemuri, B. Yezierski, and T. Mareci. Von mises-fisher mixture model of the diffusion odf. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, pages 65–68. IEEE, 2006.

[31] R. J. Muirhead. *Aspects of multivariate statistical theory*. John Wiley, 1982.

[32] C. R. Rao. *Linear Statistical Inference and its Applications*. Wiley, New York, 2nd edition, 1973.

[33] E. Rasmussen. Clustering algorithms. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, pages 419–442. Prentice Hall, New Jersey, 1992.

[34] J. Reisinger, A. Waters, B. Silverthorn, and R. J. Mooney. Spherical topic models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 903–910, 2010.

[35] S. Ryali, T. Chen, K. Supekar, and V. Menon. A parcellation scheme based on von mises-fisher distributions and markov random fields for segmenting brain regions using resting-state fmri. *NeuroImage*, 65:83–96, 2013.

[36] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley (Reading MA), 1989.

[37] G. Salton and M. J. McGill. *Introduction to Modern Retrieval*. McGraw-Hill Book Company, 1983.

[38] S. Sra. *Matrix Nearness Problems in Data Mining*. PhD thesis, Univ. of Texas at Austin, 2007.

[39] S. Sra. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$. *Computational Statistics*, Apr. 2009. Accepted.

[40] S. Sra and D. Karp. The multivariate Watson distribution: Maximum-likelihood estimation and other aspects. *Journal of Multivariate Analysis (JMVA)*, 114:256–269, 2013.

[41] J. Straub, J. Chang, O. Freifeld, and J. W. Fisher III. A dirichlet process mixture model for spherical data. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 930–938, 2015.

[42] A. Strehl and J. Ghosh. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[43] A. Tanabe, K. Fukumizu, S. Oba, T. Takenouchi, and S. Ishii. Parameter estimation for von Mises-Fisher distributions. *Computational Statistics*, 22(1):145–157, 2007.

[44] H. Tang, S. M. Chu, and T. S. Huang. Generative model-based speaker clustering via mixture of von mises-fisher distributions. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4101–4104. IEEE, 2009.

[45] D. E. Tyler. Statistical analysis for the angular central gaussian distribution on the sphere. *Biometrika*, 74(3):579–589, 1987.

[46] G. S. Watson. The statistics of orientation data. *The Journal of Geology*, pages 786–797, 1966.