Modeling data using directional distributions: Part II

Suvrit Sra Prateek Jain Inderjit Dhillon

Feb. 2007
(Revised: May 2008)

Department of Computer Sciences
The University of Texas at Austin

TR-# TR-07-05

Abstract

High-dimensional data is central to most data mining applications, and only recently has it been modeled via directional distributions. In [Banerjee et al., 2003] the authors introduced the use of the von Mises-Fisher (vMF) distribution for modeling high-dimensional directional data, particularly for text and gene expression analysis. The vMF distribution is one of the simplest directional distributions. The Watson, Bingham, and Fisher-Bingham distributions provide distributions with an increasing number of parameters and thereby commensurately increased modeling power. This report provides a followup study to the initial development in [Banerjee et al., 2003] by presenting Expectation Maximization (EM) procedures for estimating parameters of a mixture of Watson (moW) distributions. The numerical challenges associated with parameter estimation for both of these distributions are significantly more difficult than for the vMF distribution. We develop new numerical approximations for estimating the parameters permitting us to model real-life data more accurately. Our experimental results establish that for certain data sets improved modeling power translates into better results.

1 Introduction

Directional distributions provide a rich class of probabilistic models for characterizing vectorial data whose relative spatial orientations are of greater importance than their magnitude. Mardia and Jupp [2000] enlist numerous applications of directional data; however, before the work of [Banerjee et al., 2003], directional distributions had not been formally applied to modeling high-dimensional data.

We highlight at this point the fact that just as Euclidean distance is related to the multivariate Gaussian, so is cosine-similarity related to the von Mises-Fisher (vMF) distribution—the natural directional distribution on a unit hypersphere [Mardia and Jupp, 2000]. Owing to this reason, directional models seem particularly suitable for text and gene-expression data and empirical success of the cosine-similarity and Pearson correlation corroborates this belief. However, the vMF distribution can sometimes be too restrictive. In situations where data is axially symmetric, for e.g., for diametric clustering of anti-correlated genes [Dhillon et al., 2003], the vMF distribution fails to capture the inherent patterns. Indeed, as we show later in this report, the diametric clustering procedure of [Dhillon et al., 2003] is a limiting case of a mixture of Watson distributions, which are axially symmetric directional distributions. Data that exhibits several axes of symmetry can be suitably modeled with Bingham distributions that generalize the Watson distributions, though at the expense of dramatically more difficult parameter estimation.

In this report we discuss generative mixture-models based on the Watson distributions. In particular we derive an EM algorithm for performing the parameter estimation for a mixture of Watson

(moW) distributions. Akin to the vMF case [Banerjee et al., 2003], parameter estimation for the Watson distribution also turns out to be challenging as it entails solving difficult non-linear equations. We derive accurate numerical approximations for solving these nonlinear equations—these approximations are crucial for an efficient implementation.

As an outcome of our EM algorithms for doing mixture modeling with both these densities we formulate clustering procedures, which in themselves are further interesting as they provide a theoretical basis for the diametric k-means algorithm of Dhillon et al. [2003].

We would like to point out that at the time of writing this report, other work performing mixture modeling for Watson distributions has appeared in the literature [Bijral et al., 2007]. However, the authors of [Bijral et al., 2007] were not aware of the relation of mixtures of Watson distributions to diametric clustering, which we describe in this paper. The parameter estimates derived by them are also different from our approach (in fact [Bijral et al., 2007] follow the approach of [Banerjee et al., 2005] to obtain their parameter estimates).

2 Background

In this section we summarize some background material about directional distributions to provide an introduction to the uninitiated reader. Those who are familiar with these basics can straightway skip to the next section. The material in this section is based upon [Mardia and Jupp, 2000], though all the proofs are of our own construction.

Let $\mathbb{S}^{p-1}$ denote the p-dimensional unit hypersphere, i.e., $\mathbb{S}^{p-1} = \{\boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}^p, \text{and } \|\boldsymbol{x}\|_2 = 1\}$. All the densities that we describe will be defined on the surface of this unit hypersphere. We denote the probability element on $\mathbb{S}^{p-1}$ by $d\mathbb{S}^{p-1}$, and parameterize $\mathbb{S}^{p-1}$ by polar coordinates $(r, \boldsymbol{\theta})$, where $r = 1$, and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_{p-1}]$. Consequently $x_i = \sin\theta_1 \cdots \sin\theta_{i-1} \cos\theta_i$ for $1 \leq i < p$, and $x_p = \sin\theta_1 \cdots \sin\theta_{p-1}$. Given this parameterization, it is easy to show that $d\mathbb{S}^{p-1} = \big(\prod_{k=2}^{p-1} \sin^{p-k}\theta_{k-1}\big) d\boldsymbol{\theta}$.

2.1 Uniform distribution

The uniform distribution on $\mathbb{S}^{p-1}$ has its probability element equal to $c_p d\mathbb{S}^{p-1}$, where c_p is the normalization constant such that

$$\int_{\mathbb{S}^{p-1}} c_p d\mathbb{S}^{p-1} = 1.$$

Performing this simple integration we obtain

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

where $\Gamma(\cdot)$ is the well-known Gamma function [Abramowitz and Stegun, 1974].

2.2 The von Mises-Fisher distribution

A unit norm random vector $\boldsymbol{x}$ is said to have the p-dimensional von Mises-Fisher (vMF) distribution if its probability element is $c_p(\kappa) e^{\kappa \boldsymbol{\mu}^T \boldsymbol{x}} d\mathbb{S}^{p-1}$, where $\|\boldsymbol{\mu}\| = 1$ and $\kappa \geq 0$. The normalizing constant

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

where $I_s(\kappa)$ denotes the modified Bessel function of the first kind [Abramowitz and Stegun, 1974]. We note there that traditionally, researchers in directional statistics normalize the integration measure by the uniform measure, so that instead of $c_p(\kappa)$, one uses $c_p(\kappa) 2\pi^{p/2}/\Gamma(p/2)$; as far as parameter estimation is concerned, this distinction is immaterial, and we shall ignore it for the rest of this report.

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

The vMF distribution is one of the simplest distributions for directional data and it has properties analogous to those of the multi-variate Gaussian distribution for data in $\mathbb{R}^p$. For example, the maximum entropy density on $\mathbb{S}^{p-1}$ subject to the constraint that $E[\boldsymbol{x}]$ be fixed, is a vMF density (see Rao [1973, pp. 172–174] and Mardia [1975] for details).

2.3 Watson distribution

The uniform and the vMF distributions are defined over *directions*. However, sometimes the observations are *axes*, wherein the vectors $\boldsymbol{x}$ and $-\boldsymbol{x}$ are indistinguishable [Mardia and Jupp, 2000]. To model such axial data one of the simplest densities is the Watson density, whose probability element is given by $c_p(\kappa)e^{\kappa(\boldsymbol{\mu}^T\boldsymbol{x})^2}d\mathbb{S}^{p-1}$. After integrating to determine the constant we find

$$c_p(\kappa) = \frac{\Gamma(p/2)}{2\pi^{p/2}\,{}_1F_1(\frac{1}{2},\frac{p}{2},\kappa)}, \tag{2.1}$$

where ${}_1F_1$ denotes a confluent Hypergeometric function, also known as Kummer's function (see [Abramowitz and Stegun, 1974]). Due to the $e^{\kappa(\boldsymbol{\mu}^T\boldsymbol{x})^2}$ term in the Watson density, for $\kappa > 0$ the distribution tends to concentrate around $\pm\boldsymbol{\mu}$ as κ increases, whereas for $\kappa < 0$, the density concentrates around the great circle orthogonal to $\boldsymbol{\mu}$. Since $(\boldsymbol{Q}\boldsymbol{\mu}^T\boldsymbol{Q}\boldsymbol{x})^2 = (\boldsymbol{\mu}^T\boldsymbol{x})^2$ for any orthogonal matrix $\boldsymbol{Q}$, the Watson density is rotationally invariant.

2.4 Bingham distribution

There exist some axial data sets that do not exhibit rotational symmetry, as is done by Watson distributions. In such cases, one could potentially model the data using Bingham distributions. The probability element of a Bingham distribution is given by $c_p(\boldsymbol{K})e^{\boldsymbol{x}^T\boldsymbol{K}\boldsymbol{x}}$. After integrating we find that

$$c_p(\boldsymbol{K}) = \frac{\Gamma(p/2)}{2\pi^{p/2}\,{}_1F_1(\frac{1}{2},\frac{p}{2},\boldsymbol{K})}, \tag{2.2}$$

where ${}_1F_1(\cdot,\cdot,\boldsymbol{K})$ denotes the confluent Hypergeometric function of matrix argument [Muirhead, 1982]. Note that since $\boldsymbol{x}^T(\boldsymbol{K}+\delta\boldsymbol{I}_p)\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{K}\boldsymbol{x} + \delta$, the Bingham density is identifiable only up to a constant diagonal shift in $\boldsymbol{K}$. Thus one can assume $\mathrm{Tr}(\boldsymbol{K}) = 0$, or that the smallest eigenvalue of $\boldsymbol{K}$ is zero [Mardia and Jupp, 2000]. Intuitively, one can see that the eigenvalues of $\boldsymbol{K}$ determine the axes around which the data will cluster, e.g., greatest clustering will be around the axis corresponding to the leading eigenvector of $\boldsymbol{K}$.

2.5 Additional directional distributions

There exist several additional directional distributions, each with their own unique characteristics. We omit a discussion of these densities and refer the reader to Mardia and Jupp [2000] for more details. For the remainder of this report we will focus our attention on the Watson distribution, briefly touching upon the Bingham density before concluding.

Bingham-Mardia distributions Mardia and Jupp [2000] remark that 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 they suggest the density

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \kappa, \nu) = c_p(\kappa) e^{\kappa(\boldsymbol{\mu}^T \boldsymbol{x} - \nu)^2}, \tag{2.3}$$

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 [Mardia and Jupp, 2000]. The density is

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\mu}_0, \kappa, \kappa_0) = c_p(\kappa_0, \kappa, \boldsymbol{\mu}_0^T \boldsymbol{\mu}) e^{\kappa_0 \boldsymbol{\mu}_0^T \boldsymbol{x} + \kappa(\boldsymbol{\mu}^T \boldsymbol{x})^2}. \tag{2.4}$$

Fisher-Bingham This is one of the most general directional densities and is given by

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \kappa, \boldsymbol{A}) = c_p(\kappa, \boldsymbol{A}) e^{\kappa \boldsymbol{\mu}^T \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}. \tag{2.5}$$

There does not seem to exist a useful 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 $\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0}$ in (2.5).

3 Mixture Modeling

We propose to model the input data using a mixture of probability distributions, where each component of the mixture is a parameterized distribution. In traditional parametric modeling one estimates these parameters (thereby learning a generative model from the data) by assuming the observed data to be i.i.d., and maximizing the likelihood. More formally, assume that we have a dataset $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ which is modeled using a mixture of K distributions parameterized by $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K]$. Then the probability density at a single observation $\boldsymbol{x}_i$ is

$$p(\boldsymbol{x}_i | \boldsymbol{\Theta}) = \sum_{h=1}^{K} P(h) p(\boldsymbol{x}_i | \boldsymbol{\theta}_h, h),$$

where $p(\boldsymbol{x}_i | \boldsymbol{\theta}_h, h)$ is the density contributed by class h, which occurs with prior probability $P(h)$. The log-likelihood for the entire dataset $\mathcal{D}$ is given by (assuming the $\boldsymbol{x}_i$ to be i.i.d.)

$$\mathcal{L}(\mathcal{D}) = \sum_i \log\left(\sum_{h=1}^{K} P(h) p(\boldsymbol{x}_i | \boldsymbol{\theta}_h, h) \right) \tag{3.1}$$

The aim of maximum-likelihood parameter estimation is to maximize (3.1). However, maximizing (3.1) can prove to be quite difficult even for simple densities. Here is where the Expectation Maximization (EM) procedure comes to our rescue. By exploiting the concavity of the log function (3.1) can be bounded below with the aim of decoupling the terms inside the logarithm. The resultant lower-bound on the log-likelihood is easy to optimize iteratively, and the EM algorithm is guaranteed to find a locally optimal solution to (3.1) as a consequence.

3.1 The E-step

The E-step of an EM algorithm is simple, though often requiring some engineering for efficient implementation. Using a traditional hidden variable that indicates class membership or an equivalent auxiliary function technique we can easily obtain the E-step. Exploiting the concavity of the log function, from (3.1) we obtain

$$\mathcal{L}(\mathcal{D}) \geq \sum_{ih} \beta_{ih} \log \frac{\alpha_h p(\boldsymbol{x}_i | \boldsymbol{\theta}_h, h)}{\beta_{ih}}, \tag{3.2}$$

where $\beta_{ih} \geq 0$, and $\sum_h \beta_{ih} = 1$. Optimizing (3.2) over β_{ih} subject to the convexity restrictions on β_{ih} we obtain

$$\beta_{ih} = \frac{\alpha_h p(\boldsymbol{x}_i|h, \boldsymbol{\theta}_h)}{\sum_l \alpha_l p(\boldsymbol{x}_i|l, \boldsymbol{\theta}_l)}. \tag{3.3}$$

To see that this choice of β_{ih} yields the optimum value observe that for each i we essentially need to maximize $-\mathrm{KL}(\boldsymbol{\beta}_i \| \boldsymbol{\delta}_i) \leq 0$, where $\delta_{ih} \propto \alpha_h p(\boldsymbol{x}_i|\boldsymbol{\theta}_h, h)$. The EM algorithm interprets the β_{ih} values as $p(h|\boldsymbol{x}_i, \boldsymbol{\theta}_h)$.

3.1.1 Hard assignments

Traditionally, to ease computational burdens, the following *hard-assignment* heuristic (hard-clustering) is used

$$\beta_{ih} = \begin{cases} 1, & \text{if } h = \operatorname{argmax}_{h'} \log\big(\alpha_{h'} c_d(\kappa_{h'}) + \kappa_{h'} (\boldsymbol{x}_i^T \boldsymbol{\mu}_{h'})^2\big) \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

In this case, the M-step (see §3.2) also usually simplifies considerably. Since the hard-assignment heuristic is general, we do not provide specific simplifications for all the derivations below, though we will come back to certain interesting special cases. Hard-assignments maximize a lower-bound on the incomplete log-likelihood of the data, as is evident from (3.2). The fact that this lower-bound is tight is proved in [Banerjee et al., 2005].

3.2 The M-step

The main difficulty during distribution estimation for directional distributions lies in the parameter estimation in an M-step. In this step we maximize (3.2) w.r.t. the parameters $\boldsymbol{\theta}_h$, while keeping β_{ih} fixed. Formally, the M-step is

$$\max_{\boldsymbol{\Theta}} \quad \sum_{ih} \beta_{ih} \log \alpha_h p(\boldsymbol{x}_i|\theta_h), \tag{3.5}$$

subject to $\boldsymbol{\Theta} \in \Omega$, where the latter is some set (usually convex) describing the space of parameters. We assume that for $\boldsymbol{\Theta} = [\theta_1, \ldots, \theta_K]$ the individual class parameters θ_h are independent of each other. Hence, the maximization (3.5) is essentially a concatenation of K different maximization problems.

Since α_h is the prior for the h-th class, we maximize (3.5) w.r.t. α_h subject to the restriction that $\sum_h \alpha_h = 1$ to obtain

$$\alpha_h = \frac{1}{N} \sum_i \beta_{ih}. \tag{3.6}$$

The difficulty of obtaining the other parameters depends on the distribution under question. Below we list the M-step estimation for the vMF, Watson, and Bingham distributions. Other directional distributions can be handled in a similar way, but we restrict our attention to these three because we have provided fast approximate parameter estimates for these distributions.

3.2.1 M-step for Watson

For the Watson distribution we further estimate $\boldsymbol{\mu}_h$ and κ_h for all the mixture components. The maximization problem (3.5) becomes

$$\begin{aligned} \max_{\boldsymbol{\mu}_h, \kappa_h} \quad & \sum_{ih} \beta_{ih} \big(- \log {}_1F_1(\tfrac{1}{2}, \tfrac{d}{2}, \kappa_h) + \kappa_h (\boldsymbol{\mu}_h^T \boldsymbol{x}_i)^2 \big), \\ \text{subject to} \quad & \boldsymbol{\mu}_h^T \boldsymbol{\mu}_h = 1. \end{aligned} \tag{3.7}$$

The corresponding Lagrangian is

$$L(\{(\boldsymbol{\mu}_h, \kappa_h)\}) = \sum_{ih} \beta_{ih}\big(-\log {}_1F_1(\tfrac{1}{2}, \tfrac{d}{2}, \kappa_h) + \kappa_h(\boldsymbol{\mu}_h^T \boldsymbol{x}_i)^2\big) + \sum_h \lambda_h(\boldsymbol{\mu}_h^T \boldsymbol{\mu}_h - 1). \tag{3.8}$$

Computing $\partial L/\partial \boldsymbol{\mu}_h$ and $\partial L/\partial \kappa_h$, setting them to zero, and enforcing the normalization constraints on $\boldsymbol{\mu}_h$ we obtain the parameter estimates

$$\boldsymbol{\mu}_h = \frac{\sum_i \beta_{ih} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\mu}_h}{\|\sum_i \beta_{ih} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\mu}_h\|} = \frac{\boldsymbol{A}_h \boldsymbol{\mu}_h}{\|\boldsymbol{A}_h \boldsymbol{\mu}_h\|}, \tag{3.9}$$

$$\kappa_h = \text{Solve}\left[\frac{{}_1F_1'(\frac{1}{2}, \frac{d}{2}, \kappa_h)}{{}_1F_1(\frac{1}{2}, \frac{d}{2}, \kappa_h)} = \frac{\sum_i \beta_{ih} (\boldsymbol{x}_i^T \boldsymbol{\mu}_h)^2}{\sum_i \beta_{ih}}\right]. \tag{3.10}$$

Observe that in (3.9) the variable $\boldsymbol{\mu}_h$ occurs on both sides of the equation, thereby necessitating an iterative solution. From (3.9) we see that $\boldsymbol{\mu}_h$ is given by the leading left-singular vector of the matrix $\boldsymbol{A}_h$. Obtaining κ_h requires the solution of the nonlinear equation in (3.10), as indicated by the Solve[·] operation. However, for high-dimensionality (large d), a nonlinear root-finder for solving (3.10) can be very time consuming and somewhat of an engineering challenge due to numerical issues (such as overflow owing to the huge magnitude that ${}_1F_1$ might attain). In Section 4 we derive an asymptotic approximation for computing κ_h which is extremely efficient.

3.3 Algorithms for moW

The clustering algorithms for moW distributions are based on soft and hard-assignment schemes and are titled soft-moW and hard-moW respectively. The soft-moW algorithm (Algorithm 3.1) estimates the parameters of the mixture model exactly following the derivations in Section 3.2 using EM. Hence, it assigns soft (or probabilistic) labels to each point that are given by the posterior probabilities of the components of the mixture conditioned on the point. On termination, the algorithm gives the parameters $\Theta = \{\alpha_h, \boldsymbol{\mu}_h, \kappa_h\}_{h=1}^K$ of the K Watson distributions that model the dataset $\mathcal{X}$, as well as the *soft-clustering*, i.e., the posterior probabilities $p(h|\boldsymbol{x}_i, \Theta)$, for all h and i (given by the β_{ih} values)

Algorithm 3.1: soft-moW

SOFTMOW($\mathcal{X}$)
Input: $\mathcal{X} \in \mathbb{S}^{p-1}$, K: number of clusters
Output: Soft clustering of $\mathcal{X}$ over a mixture of K Watson distributions
{Initialize}
$\alpha_h, \boldsymbol{\mu}_h, \kappa_h$ for $1 \leq h \leq K$
while not converged
  The E (Expectation) step of EM
  for $i = 1$ **to** N
    for $h = 1$ **to** K
      $p_h(\boldsymbol{x}_i|\boldsymbol{\theta}_h) \leftarrow c_p(\kappa_h) e^{\kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x}_i}$
      $\beta_{ih} = p(h|\boldsymbol{x}_i, \boldsymbol{\Theta}) \leftarrow \dfrac{\alpha_h p_h(\boldsymbol{x}_i|\boldsymbol{\theta}_h)}{\sum_{l=1}^{K} \alpha_l p_l(\boldsymbol{x}_i|\boldsymbol{\theta}_l)}$
    end *for.*
  end *for.*
  The M (Maximization) step of EM
  for $h = 1$ **to** K
    $\alpha_h \leftarrow \frac{1}{N} \sum_{i=1}^{N} \beta_{ih}$
    $\boldsymbol{\mu}_h \leftarrow \frac{\sum_i \beta_{ih} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\mu}_h}{\| \sum_i \beta_{ih} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\mu}_h \|}$; $\quad \bar{r}_h = \frac{\sum_i \beta_{ih} (\boldsymbol{x}_i^T \boldsymbol{\mu}_h)^2}{N \alpha_h}$
    $\kappa_h \leftarrow \text{Solve}\Big[\frac{{}_1F_1'(\frac{1}{2}, \frac{p}{2}, \kappa_h)}{{}_1F_1(\frac{1}{2}, \frac{p}{2}, \kappa_h)} = \bar{r}_h \Big]$
  end *for.*
end *while.*

The hard-moW algorithm (Algorithm 3.2) estimates the parameters of the mixture model using a hard assignment. In other words, we do the assignment of the points based on a derived posterior distribution, wherein the E-step that estimates β_{ih} is replaced by

$$\beta_{ih} \leftarrow \begin{cases} 1, & \text{if } \; h = \underset{h'}{\text{argmax}} \; \alpha_{h'} \; p_{h'}(\boldsymbol{x}_i|\boldsymbol{\theta}_{h'}) \\ 0, & \text{otherwise.} \end{cases} \tag{3.11}$$

After these hard assignments each point $\boldsymbol{x}_i$, *belongs* to a single cluster. Upon termination, Algorithm 3.2 yields a hard clustering of the data and the parameters $\boldsymbol{\Theta} = \{\alpha_h, \boldsymbol{\mu}_h, \kappa_h\}_{h=1}^{K}$.

3.4 Relation to diametric clustering

The Diametric Clustering algorithm of Dhillon et al. [2003] groups together both correlated and anti-correlated data points. This amounts to grouping points while respecting axial symmetry. Immediately one might ask the question whether the diametric clustering procedure bears a relation to clustering based on mixtures of distributions that respect axial symmetry, i.e., distributions that essentially treat $\pm\boldsymbol{x}$ as the same. We answer this question in the affirmative, and show that the diametric clustering procedure of Dhillon et al. [2003] is a limiting case of EM for a mixture of Watson distributions. To that end, first we recapitulate the diametric clustering procedure (taken from [Dhillon et al., 2003]) as Algorithm 3.3.

Algorithm 3.2: hard-moW

HARDMOW($\mathcal{X}$)
Input: $\mathcal{X} \in \mathbb{S}^{p-1}$, K: number of clusters
Output: A disjoint K-partitioning of $\mathcal{X}$
{Initialize}
α_h, $\boldsymbol{\mu}_h$, κ_h for all $1 \leq h \leq K$
while not converged
  {The Hardened E-step of EM}
  for $i = 1$ **to** N
    for $h = 1$ **to** K
      $p_h(\boldsymbol{x}_i|\boldsymbol{\theta}_h) \leftarrow c_p(\kappa_h)e^{\kappa_h \boldsymbol{\mu}_h^T \boldsymbol{x}_i}$
      $\beta_{ih} \leftarrow \begin{cases} 1, & \text{if } \ h = \text{argmax}_{h'} \ \ p_{h'}(\boldsymbol{x}_i|\boldsymbol{\theta}_{h'}) \\ 0, & \text{otherwise.} \end{cases}$
    end *for*
  end *for*
  {The M-step of EM}
  Same as in Algorithm 3.1
end *while*.

Algorithm 3.3: diametric K-means

DIAMETRIC($\mathcal{X}$, K)
Input: $\mathcal{X} \in \mathbb{S}^{p-1}$
Output: A disjoint K-partitioning $\mathcal{X}_k$ of $\mathcal{X}$
{Initialize}
$\boldsymbol{\mu}_h$ for $1 \leq h \leq K$
while not converged
  {The E-step of EM}
  Set $\mathcal{X}_h \leftarrow \emptyset$ for all $1 \leq h \leq K$
  for $i = 1$ **to** N
    $\mathcal{X}_h \leftarrow \mathcal{X}_h \cup \{\boldsymbol{x}_i\}$ where $h = \text{argmax}_{h'}(\boldsymbol{x}_i^T \boldsymbol{\mu}_{h'})^2$
  end *for*
  {The M (Maximization) step of EM}
  for $h = 1$ **to** K
    $\boldsymbol{K}_h = [\boldsymbol{x}_i]$ such that $\boldsymbol{x}_i \in \mathcal{X}_h$
    $\boldsymbol{\mu}_h \leftarrow \frac{\boldsymbol{K}_h \boldsymbol{\mu}_h}{\|\boldsymbol{K}_h \boldsymbol{\mu}_h\|}$
  end *for*
end*while*.

From Algorithm 3.3 it is evident how it may be derived as a limiting case of EM for a mixture of Watson distributions. The first view is based on a limiting view of soft-moW as $\kappa_h \to \infty$, because this sends $\beta_{ih} \to \{0, 1\}$. The second limiting view comes from ignoring kappas in hard-moW, by setting all of them to some fixed value κ^*. We omit the details for brevity as they are analogous to the reduction in [Banerjee et al., 2005].

4 Approximating κ

In this section we exploit some of the properties of the confluent hypergeometric function ${}_1F_1$ to obtain an extremely efficient approximation to (3.10). It is well known [Abramowitz and Stegun,

1974] that

$$\frac{\partial\, {}_1F_1(a,b,z)}{\partial z} = \frac{a}{b}\, {}_1F_1(a+1,b+1,z). \tag{4.1}$$

Assuming that b is relatively large we approximate (4.1) to write

$$\frac{a}{b}\, {}_1F_1(a+1,b+1,z) \approx \frac{a}{b-1}\, {}_1F_1(a+1,b,z), \tag{4.2}$$

essentially replacing b by $b-1$. A useful identity that ${}_1F_1$ satisfies is

$$(a+z)\, {}_1F_1(a+1,b,z) + (b-a-1)\, {}_1F_1(a,b,z) + (1-b)\, {}_1F_1(a+1,b-1,z) = 0. \tag{4.3}$$

We approximate the last term in the identity above by ${}_1F_1(a+1,b,z)$. Hence we get the new approximation

$$(a+b-1+z)\, {}_1F_1(a+1,b,z) \approx (a+b-1)\, {}_1F_1(a,b,z). \tag{4.4}$$

Recall that in (3.10) we needed to essentially solve

$$\frac{{}_1F_1'(a,b,z)}{{}_1F_1 a,b,z} = \bar{r},$$

where a, b, z, and $\bar{r}$ are defined appropriately. Using (4.2) in (4.4) we obtain

$$(a+b-1+z)\frac{b-1}{a}\, {}_1F_1'(a,b,z) \approx (a+b-1)\, {}_1F_1(a,b,z). \tag{4.5}$$

We solve this latter approximation (writing ${}_1F_1'(a,b,z)/\, {}_1F_1(a,b,z) = \bar{r}$) to obtain

$$z \approx \frac{a(a+b-1)}{(b-1)\bar{r}}. \tag{4.6}$$

However, in practice we have observed that the "corrected"-approximation

$$z \approx (a+b-1)\left(\frac{1}{1-\bar{r}} - \frac{a}{(b-1)\bar{r}}\right), \tag{4.7}$$

leads to much better accuracy. This accuracy may be viewed as the result of incorporating the relative error term

$$\epsilon = \frac{{}_1F_1'(a,b,z)}{{}_1F_1'(a,b,z) - {}_1F_1(a,b,z)},$$

into (4.5), so that we solve the "corrected"-approximation

$$((a+b-1)(1+\epsilon)+z)\frac{b-1}{a}\, {}_1F_1'(a,b,z) \approx (a+b-1)\, {}_1F_1(a,b,z). \tag{4.8}$$

This solution of (4.8) is given by (4.7), and it yields significantly better accuracy than (4.6) in practice.

4.1 A more careful look at the approximations

It is obvious that the error of approximation depends heavily upon the parameters a, b and z. Depending upon these parameters, we have the following four approximations (all of these are variations of (4.7)).

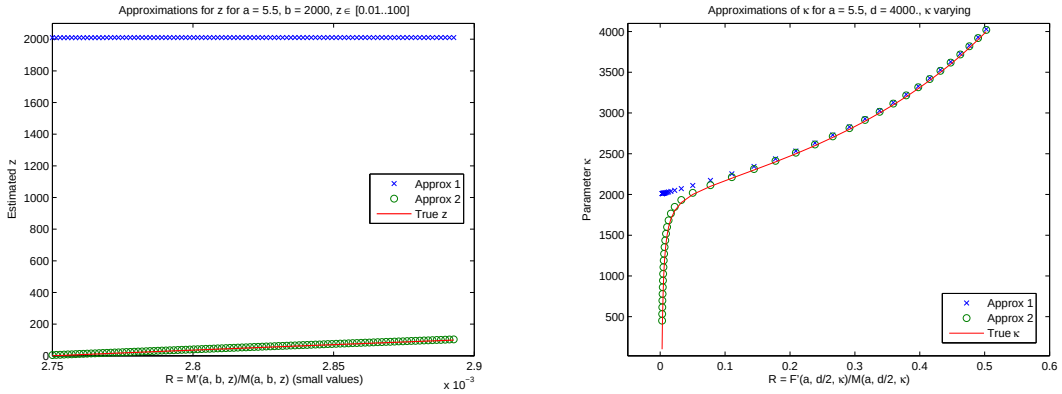
$$z \approx (a+b-1)\frac{1}{1-\bar{r}} \tag{A1}$$

$$z \approx (a+b-1)\left(\frac{1}{1-\bar{r}} - \frac{a}{(b-1)\bar{r}}\right) \tag{A2}$$

$$z \approx (a+b-1)\left(\frac{1}{1-\bar{r}} + \frac{a-1}{(b-1)\bar{r}}\right) \tag{A3}$$

$$z \approx (a+b-1)\left(\frac{1}{1-\bar{r}} - \frac{a}{b\bar{r}}\right). \tag{A4}$$

We conducted some experiments to determine the parameter ranges for which these approximations work well. Figure 1 displays the behavior of the approximations (A1) and (A2) across a range of z as a and b are held fixed. We display $\bar{r} = {}_1F_1'/{}_1F_1$ on the X-axis, since the approximations are functions of $\bar{r}$, and show varying degrees of accuracy for small or large values of $\bar{r}$.



(a) (b)

Figure 1: Approximation for varying z with parameters a and b are held fixed at 5.5, and 2000, respectively. Subfigure (a) compares (A1) and (A2) for small values of $z \in [0.01..100]$. Subfigure (b) compares (A1) and (A2) for larger values of $z \in [100..4000]$. Notice that as $\bar{r} = {}_1F_1'/{}_1F_1$ increases, both approximations become accurate for large values of z. In fact, (A1) is more accurate than (A2) for larger values of z

Figure 2 reports results similar to those in Figure 1 except that now $a = 0.5$ is used, whereby to attain better approximations we had to use (A3). In this case (A2) leads to very poor approximations.

Figure 3 shows approximations A1 and A3 for $a = 0.5$, as b is varied and z is held fixed. From the results above it may seem that approximations (A3) and (A4) perform similarly. A small attestation to this observation is provided by Figure 4 below.

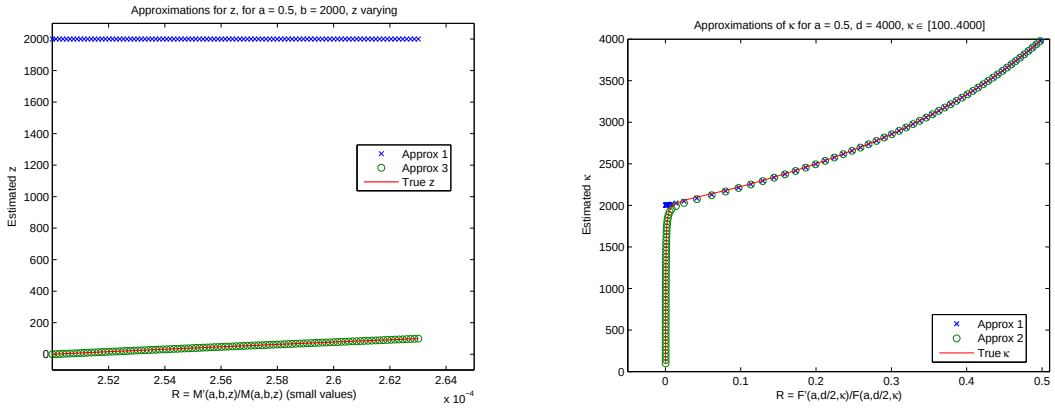
5 Discussion and Future Work

In this report we presented simple EM procedures for doing parameter estimation for a mixture of Watson distributions. We presented simple and efficient numerical estimates for solving the transcendental equations that arise while performing the M-step for parameter estimation. We also showed empirical verification to exhibit the accuracy of our estimates. Additionally, we also showed how the diametric clustering algorithm of Dhillon et al. [2003] may be obtained as a limiting case of an EM procedure for moW distributions.

An extension of the EM method to data modeled using a mixture of Bingham distributions remains a part of our future work. The parameter estimation for this case is significantly more challenging than for the moW case, and remains the greatest barrier to the development of an efficient EM method for mixtures of Bingham distributions.

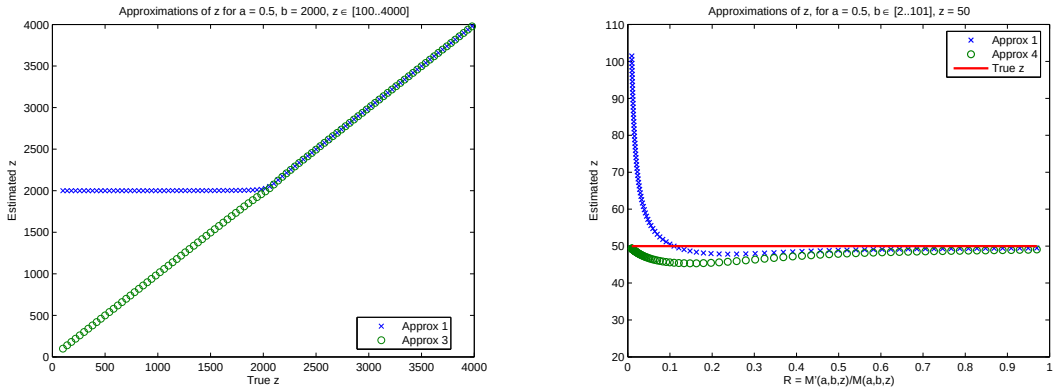
References

M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables*. Dover Publ. Inc., New York, June 1974. ISBN 0486612724.



(a) (b)

Figure 2: Approximation for varying z with with parameters a and b are held fixed at 0.5, and 2000.0, respectively. Subfigure (a) compares (A1) and (A2) for small values of $z \in [0.01..100]$. Subfigure (b) compares (A1) and (A2) for larger values of $z \in [100..4000]$. Notice that as $\bar{r} = {}_1F_1'/{}_1F_1$ increases, both approximations become accurate for large values of z. In fact, (A1) is more accurate than (A2) for larger values of z.



(a) (b)

Figure 3: Subfigure (a) shows the approximation of Figure 2(b) but with the true value of z as the x-axis. Subfigure (b) shows a plot of how approximation (A1) compares against (A4) as b is varied from 2 to 101. The true z was held constant at 50, and $a = 0.5$ was also fixed. We see that (A4) consistently outperforms (A1).

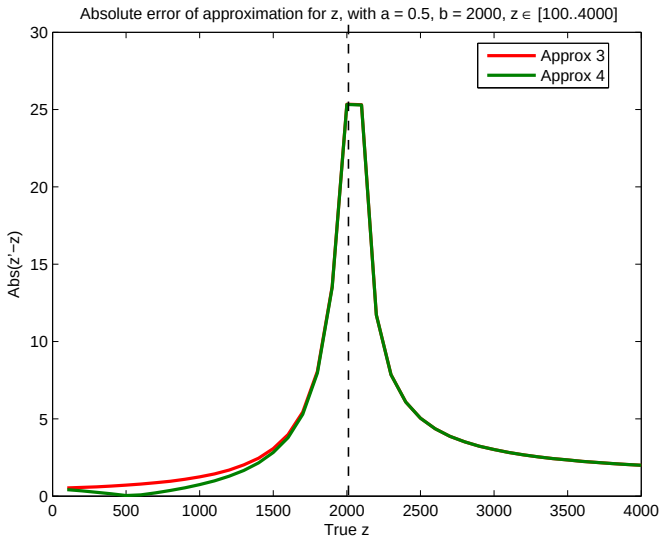


Figure 4: Absolute error of approximation of z for fixed $a = 0.5$, $b = 2000$, and $z \in [100..4000]$. After a point both (A3) and (A4) yield the same results. Note from the figure how the error shoots up when the true value of z is close to b. It is a known fact that asymptotic approximations for Hypergeometric functions break down when a is small, but both b and z are of comparable magnitude. Hence, some further scope of improvement is possible.

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Expectation maximization for clustering on hyperspheres. Technical Report TR-03-07, Department of Computer Sciences, University of Texas, February 2003.

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6:1345–1382, Sep 2005.

A. Bijral, M. Breitenbach, and G. Z. Grudic. Mixture of Watson Distributions: A Generative Model for Hyperspherical Embeddings. In *AISTATS*, 2007.

I. S. Dhillon and S. Sra. Modeling data using directional distributions. Technical Report TR-03-06, Computer Sciences, The Univ. of Texas at Austin, January 2003.

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

R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison Wesley, 1998.

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

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

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

K. E. Muller. Computing the confluent hypergeometric function, $m(a, b, x)$. *Numerische Mathematik*, 90(1):179–196, 2001.

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

A Mathematical Background

For completeness, we include some of the relevant mathematical background in this appendix. We recommend reading the appendix of our first technical report on directional distributions [Dhillon and Sra, 2003] for the mathematical background necessary for some of the derivations in this appendix (e.g., the $\Gamma(x)$ function, the $\sin^n(x)$ integral etc.).

A.1 Hypergeometric functions

Hypergeometric functions provide one of the richest classes of functions in analysis. Traditionally the name "hypergeometric function" is used for Gauss' hypergeometric function

$$ {}_2F_1(a, b; c; z) = \sum_{k \geq 0} \frac{a^{\overline{k}} b^{\overline{k}}}{c^{\overline{k}} k!} z^k, \tag{A.1} $$

where $a^{\overline{k}}$ denotes the rising factorial (notation adopted from [Graham et al., 1998]), which is also denoted by the Pochhammer symbol $(a)_k = a(a+1)\dots(a+k-1)$. The generalized hypergeometric function ${}_pF_q$ is defined analogously as

$$ {}_pF_q = F(a_1, \dots, a_p; b_1, \dots, b_q; z) = \sum_{k \geq 0} \frac{a_1^{\overline{k}} \cdots a_p^{\overline{k}}}{b_1^{\overline{k}} \cdots b_q^{\overline{k}} k!} z^k. \tag{A.2} $$

The reader is referred to [Abramowitz and Stegun, 1974, Graham et al., 1998] for more information on Hypergeometric functions. Several online resources also provide useful information.

For directional distributions, the hypergeometric function of interest is the confluent hypergeometric ${}_1F_1(a, b, z)$, also called *Kummer's* function. We now prove the two identities (4.1) and (4.3) that proved crucial to the derivation of our approximations.

Lemma A.1 (Derivative of ${}_1F_1$)**.**

$$ \frac{d^n}{dz^n}\big({}_1F_1(a, b, z)\big) = \frac{a^{\overline{n}}}{b^{\overline{n}}} {}_1F_1(a+n, b+n, z). $$

Proof. We prove that $(d/dz)\,{}_1F_1(a, b, z) = (a/b)\,{}_1F_1(a+1, b+1, z)$, and the remainder of the proof follows by induction. We have

$$ \begin{aligned} \frac{d}{dz}\,{}_1F_1(a, b, z) &= \sum_{k \geq 1} \frac{a^{\overline{k}}}{b^{\overline{k}}(k-1)!} z^{k-1} \\ &= \frac{a}{b} \sum_{k \geq 1} \frac{(a+1)^{\overline{k-1}}}{(b+1)^{\overline{k-1}}(k-1)!} z^{k-1} \\ &= \frac{a}{b} \sum_{k \geq 0} \frac{(a+1)^{\overline{k}}}{(b+1)^{\overline{k}} k!} z^k. \end{aligned} $$

□

Lemma A.2. *The following identity holds.*

$$ (a+z)\,{}_1F_1(a+1, b, z) + (b-a-1)\,{}_1F_1(a, b, z) + (1-b)\,{}_1F_1(a+1, b-1, z) = 0. $$

Proof. The sum of the coefficients of $\frac{a^{\overline{k}}}{b^{\overline{k}} k!} z^k$ from each of the three terms above is

$$ (a+k) + \frac{k(b+k-1)}{a} + b - (a+1) - \frac{(a+k)(b+k-1)}{a} = 0. $$

For obtaining the above sum of coefficients we used the easily proved identity $x^{\overline{m+n}} = x^{\overline{m}}(x+m)^{\overline{n}}$ (from which results such as $x^{\overline{k-1}} = x^{\overline{k}}/(x+k-1)$ trivially follow). □

Lemma A.3. *Given that $a > 0$ and $b > a$, the following identity holds.*

$$\frac{\Gamma(b-a)\Gamma(a)}{\Gamma(b)}\,{}_1F_1(a,b,z) = \int_0^1 e^{zt}t^{a-1}(1-t)^{b-a-1}dt, \tag{A.3}$$

Proof. Expand e^{zt} in its power series and integrate term by term to obtain the answer. We use the fact that (see [Abramowitz and Stegun, 1974] or [Dhillon and Sra, 2003] for basic facts about the $\Gamma(x)$ function)

$$\int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

along with the simple relation $a^{\overline{k}} = \Gamma(a+k)/\Gamma(a)$. □

A.2 Normalization constant for the Watson distribution

Now we derive the normalization constant for the Watson distribution over the unit hypersphere $\mathbb{S}^{p-1}$. The Watson density may be written as

$$W_d(\kappa, \boldsymbol{\mu}; \boldsymbol{x}) = c_p(\kappa)e^{\kappa(\boldsymbol{\mu}^T\boldsymbol{x})^2}, \quad \boldsymbol{x} \in \mathbb{S}^{p-1}, \kappa \in \mathbb{R}. \tag{A.4}$$

Normally, in the statistics literature this constant is computed relative to the uniform measure, which amounts to normalizing it by the volume of the unit hypersphere. Now, we make a change of variables to polar coordinates and integrate (A.4) over $\mathbb{S}^{p-1}$. Thus,

$$\begin{aligned}\int_{\mathbb{S}^{p-1}} W_d(\kappa, \boldsymbol{\mu}; \boldsymbol{x})d\boldsymbol{x} &= \int_{\mathbb{S}^{p-1}} c_p(\kappa)e^{\kappa(\boldsymbol{\mu}^T\boldsymbol{x})^2} = 1 \\ &= \int_0^{2\pi} d\theta_{p-1}\int_0^{\pi} e^{\kappa\cos^2\theta_1}\sin^{p-2}\theta_1 d\theta_1 \prod_{j=3}^{p-1}\int_0^{\pi}\sin^{p-j}\theta_{j-1}d\theta_{j-1} \\ &= 2\pi \times I \times \pi^{\frac{p-3}{2}}\frac{1}{\Gamma(\frac{p-1}{2})}.\end{aligned} \tag{A.5}$$

We now look at the integral denoted by I in the last step above. Let $\phi \leftarrow \theta - \frac{\pi}{2}$. Then we have

$$I = \int_0^{\pi} e^{\kappa\cos^2\theta_1}\sin^{p-2}\theta_1 d\theta_1 = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{\kappa\sin^2\theta_1}\cos^{p-2}\theta_1 d\theta_1.$$

The integrand above is an even function, hence we have

$$I = 2\int_0^{\frac{\pi}{2}} e^{\kappa\sin^2\theta_1}\cos^{p-2}\theta_1 d\theta_1.$$

Making the variable substitution $t \leftarrow \sin^2(\theta_1)$ and using (A.3), we can rewrite I as

$$I = \int_0^1 e^{\kappa t}t^{-\frac{1}{2}}(1-t)^{\frac{p}{2}-\frac{1}{2}-1}dt = \frac{\Gamma(\frac{p}{2}-\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{p}{2})}\,{}_1F_1(\frac{1}{2},\frac{p}{2},\kappa).$$

Hence (A.5) becomes

$$2\pi^{\frac{p-1}{2}}\frac{\Gamma(1/2)}{\Gamma(p/2)}\,{}_1F_1(\frac{1}{2},\frac{p}{2},\kappa).$$

Hence $c_p(\kappa)$ is (using the fact that $\Gamma(1/2) = \sqrt{\pi}$)

$$\frac{\Gamma(p/2)}{2\pi^{p/2}}\, {}_1F_1(\frac{1}{2}, \frac{p}{2}, \kappa)^{-1}. \tag{A.6}$$

Normalizing (A.6) by the surface area of the unit hypersphere, i.e., by $\Gamma(p/2)/(2\pi^{p/2})$ (see [Dhillon and Sra, 2003]) one obtains

$$c_p(\kappa) = \frac{1}{{}_1F_1(\frac{1}{2}, \frac{p}{2}, \kappa)}. \tag{A.7}$$

A.3 Computing ${}_1F_1$

The confluent hypergeometric function ${}_1F_1(a, b, z)$ appears to be fairly simple. However, for the range of arguments expected when dealing with high-dimensional distributions it can be difficult to efficiently compute it.

Muller [2001] discusses several different algorithms for computing ${}_1F_1$. However, the simplest of all is a simple truncated power-series. Using a multi-precision floating point computation library (such as MPFR or NTL) Algorithm A.4 can be implemented to efficiently compute ${}_1F_1$. The power-series does not always converge fast, so could additionally use Aitken's process or some other series convergence acceleration method if needed.

Algorithm A.4: Computing ${}_1F_1(a, b, z)$ via truncated series

KummerSeries(a, b, z)
Input: a, b, z: positive real numbers; τ: tolerance param.
Output: $M \approx {}_1F_1(a, b, z)$
{Initialize}
$M \leftarrow 1.0$, $R \leftarrow 1.0$
while *not converged*
  $\beta \leftarrow \frac{(a+i)*z}{(b+i)*(i+1)}$
  $R \leftarrow \beta * R$
  $M \leftarrow M + R$
  if $\beta < \tau$ **or** $M/R < \tau$
    converged $\leftarrow$ *true*
  end *if*
end *while*
return (M).