Journal of Multivariate Analysis 114 (2013) 256–269

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

The multivariate Watson distribution: Maximum-likelihood estimation and other aspects

Suvrit Sra [a,*], Dmitrii Karp [b]

[a] Max Planck Institute for Intelligent Systems, Tübingen, Germany

[b] Far Eastern Federal University, Chair of Business Informatics, 19 Okeansky prospekt, Vladivostok, 690950, Russian Federation

ARTICLE INFO

Article history:
Received 20 September 2011
Available online 7 September 2012

AMS subject classifications:
62F10
62H11
62H30
65H99
33C15

Keywords:
Watson distribution
Kummer function
Confluent hypergeometric
Directional statistics
Diametrical clustering
Special function
Hypergeometric identities

ABSTRACT

This paper studies fundamental aspects of modelling data using multivariate Watson distributions. Although these distributions are natural for modelling axially symmetric data (i.e., unit vectors where $\pm\boldsymbol{x}$ are equivalent), for high-dimensions using them can be difficult—largely because for Watson distributions even basic tasks such as maximum-likelihood are numerically challenging. To tackle the numerical difficulties some approximations have been derived. But these are either grossly inaccurate in high-dimensions [K.V. Mardia, P. Jupp, Directional Statistics, second ed., John Wiley & Sons, 2000] or when reasonably accurate [A. Bijral, M. Breitenbach, G.Z. Grudic, Mixture of Watson distributions: a generative model for hyperspherical embeddings, in: Artificial Intelligence and Statistics, AISTATS 2007, 2007, pp. 35–42], they lack theoretical justification. We derive new approximations to the maximum-likelihood estimates; our approximations are theoretically well-defined, numerically accurate, and easy to compute. We build on our parameter estimation and discuss mixture-modelling with Watson distributions; here we uncover a hitherto unknown connection to the "diametrical clustering" algorithm of Dhillon et al. [I.S. Dhillon, E.M. Marcotte, U. Roshan, Diametrical clustering for identifying anticorrelated gene clusters, Bioinformatics 19 (13) (2003) 1612–1619].

1. Introduction

Life on the surface of the unit hypersphere is more twisted than you might imagine: designing elegant probabilistic models is easy but using them is often not. This difficulty usually stems from the complicated normalising constants associated with directional distributions. Nevertheless, owing to their powerful modelling capabilities, distributions on hyperspheres continue finding numerous applications—see e.g., the excellent book *Directional Statistics* [16].

A fundamental directional distribution is the von Mises–Fisher (vMF) distribution, which models data concentrated around a mean-direction. But for data that have additional structure, vMF can be inappropriate: in particular, for axially symmetric data it is more natural to prefer the (Dimroth–Scheidegger)–Watson distribution [16,21]. And this distribution is the focus of our paper.

Three main reasons motivate our study of the multivariate Watson (mW) distribution, namely: (i) is fundamental to directional statistics; (ii) it has not received much attention in modern data-analysis setups involving high-dimensional data; and (iii) it provides a theoretical basis to "diametrical clustering", a procedure developed for gene-expression analysis [7].

Somewhat surprisingly, for high-dimensional settings, the mW distribution seems to be fairly under-studied. One reason might be that the traditional domains of directional statistics are low-dimensional, e.g., circles or spheres. Moreover, in

* Corresponding author.
E-mail addresses: suvrit@gmail.com, suvrit@tuebingen.mpg.de (S. Sra), dimkrp@gmail.com (D. Karp).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.08.010

low-dimensions numerical difficulties that are rife in high-dimensions are not so pronounced. This paper contributes theoretically and numerically to the study of the mW distribution. We hope that these contributions and the connections we make to established applications help promote wider use of the mW distribution.

1.1. Related work

Beyond their use in typical applications of directional statistics [16], directional distributions gained renewed attention in data-mining, where the vMF distribution was first used by Banerjee et al., [2,3], who also derived some *ad-hoc* parameter estimates; Non *ad-hoc* parameter estimates for the vMF case were obtained by Tanabe et al. [20].

More recently, the Watson distribution was considered in [4] and also in [18]. Bijral et al. [4] used an approach similar to that of [2] to obtain a useful but *ad-hoc* approximation to the maximum-likelihood estimates. We eliminate the *ad-hoc* approach and formally derive tight, two-sided bounds which lead to parameter approximations that are accurate and efficiently computed.

Our derivations are based on carefully exploiting properties (several *new* ones are derived in this paper) of the confluent hypergeometric function, which arises as a part of the normalisation constant. Consequently, a large body of classical work on special functions is related to our paper. But to avoid detracting from the main message and due to space limitations, we relegate highly technical details to the Appendix and to an extended version of this paper [19].

Another line of related work is based on mixture-modelling with directional distributions, especially for high-dimensional datasets. In [3], mixture-modelling using the Expectation Maximisation (EM) algorithm for mixtures of vMFs was related to cosine-similarity based K-means clustering. Specifically, Banerjee et al. [3] showed how the cosine based K-means algorithm may be viewed as a limiting case of the EM algorithm for mixtures of vMFs. Similarly, we investigate mixture-modelling using Watson distributions, and connect a limiting case of the corresponding EM procedure to a clustering algorithm called "diametrical clustering" [7]. Our viewpoint provides a new interpretation of the (discriminative) diametrical clustering algorithm and also lends generative semantics to it. Consequently, using a mixture of Watson distributions we also obtain a clustering procedure that can provide better clustering results than plain diametrical clustering alone.

2. Background

Let $\mathbb{S}^{p-1} = \{\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^p,\ \|\boldsymbol{x}\|_2 = 1\}$ be the $(p-1)$-dimensional unit hypersphere centred at the origin. We focus on axially symmetric vectors, i.e., $\pm\boldsymbol{x} \in \mathbb{S}^{p-1}$ are equivalent; this is also denoted by $\boldsymbol{x} \in \mathbb{P}^{p-1}$, where $\mathbb{P}^{p-1}$ is the projective hyperplane of dimension $p-1$. A natural choice for modelling such data is the multivariate Watson distribution [16]. This distribution is parametrised by a *mean-direction* $\boldsymbol{\mu} \in \mathbb{P}^{p-1}$, and a *concentration* parameter $\kappa \in \mathbb{R}$; its probability density function is

$$W_p(\boldsymbol{x};\, \boldsymbol{\mu}, \kappa) = c_p(\kappa) e^{\kappa(\boldsymbol{\mu}^\top \boldsymbol{x})^2}, \quad \boldsymbol{x} \in \mathbb{P}^{p-1}. \tag{2.1}$$

The normalisation constant $c_p(\kappa)$ in (2.1) is given by

$$c_p(\kappa) = \frac{\Gamma(p/2)}{2\pi^{p/2} M\left(\frac{1}{2}, \frac{p}{2}, \kappa\right)}, \tag{2.2}$$

where M is the Kummer confluent hypergeometric function defined as [8, formula 6.1(1)] or [1, formula (2.1.2)]

$$M(a, c, \kappa) = \sum_{j \geq 0} \frac{a^{\bar{j}}}{c^{\bar{j}}} \frac{\kappa^j}{j!}, \quad a, c, \kappa \in \mathbb{R}, \tag{2.3}$$

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

Observe that for $\kappa > 0$, the density concentrates around $\boldsymbol{\mu}$ as κ increases, whereas for $\kappa < 0$, it concentrates around the great circle orthogonal to $\boldsymbol{\mu}$. Observe that $(\mathbf{Q}\boldsymbol{\mu})^\top \mathbf{Q}\boldsymbol{x} = \boldsymbol{\mu}^\top \boldsymbol{x}$ for any orthogonal matrix $\mathbf{Q}$. In particular for $\mathbf{Q}\boldsymbol{\mu} = \boldsymbol{\mu}$, $\boldsymbol{\mu}^\top(\mathbf{Q}\boldsymbol{x}) = \boldsymbol{\mu}^\top \boldsymbol{x}$; thus, the Watson density is rotationally symmetric about $\boldsymbol{\mu}$.

2.1. Maximum likelihood estimation

We now consider the basic and apparently simple task of maximum-likelihood parameter estimation for mW distributions: this task turns out to be surprisingly difficult.

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{P}^{p-1}$ be i.i.d. points drawn from $W_p(\boldsymbol{x};\, \boldsymbol{\mu}, \kappa)$, the Watson density with mean $\boldsymbol{\mu}$ and concentration κ. The corresponding log-likelihood is

$$\ell(\boldsymbol{\mu}, \kappa; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = n\big(\kappa \boldsymbol{\mu}^\top \mathbf{S} \boldsymbol{\mu} - \ln M(1/2, p/2, \kappa) + \gamma\big), \tag{2.4}$$

where $\mathbf{S} = n^{-1} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top$ is the sample *scatter matrix*, and γ is a constant term that we can ignore. Maximising (2.4) leads to the following parameter estimates [16, Section 10.3.2] for the mean vector

$$\hat{\boldsymbol{\mu}} = \boldsymbol{s}_1 \quad \text{if } \hat{\kappa} > 0, \qquad \hat{\boldsymbol{\mu}} = \boldsymbol{s}_p \quad \text{if } \hat{\kappa} < 0, \tag{2.5}$$

where $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_p$ are normalised eigenvectors ($\in \mathbb{P}^{p-1}$) of the scatter matrix $\boldsymbol{S}$ corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. The concentration estimate $\hat{\kappa}$ is obtained by solving (to be more precise, 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):

$$g\left(\frac{1}{2}, \frac{p}{2}; \hat{\kappa}\right) := \frac{M'\left(\frac{1}{2}, \frac{p}{2}, \hat{\kappa}\right)}{M\left(\frac{1}{2}, \frac{p}{2}, \hat{\kappa}\right)} = \hat{\boldsymbol{\mu}}^\top \boldsymbol{S} \hat{\boldsymbol{\mu}} := r \quad (0 \leq r \leq 1), \tag{2.6}$$

where M' denotes the derivative with respect to $\hat{\kappa}$. Notice that (2.5) and (2.6) are coupled — so we need some way to decide whether to solve $g(1/2, p/2; \hat{\kappa}) = \lambda_1$ or to solve $g(1/2, p/2; \hat{\kappa}) = \lambda_p$ instead. An easy choice is to solve both equations, and select the solution that yields a higher log-likelihood. Solving these equations is much harder.

One could solve (2.6) using a root-finding method (e.g. Newton–Raphson). But, the situation is not that simple. For reasons that will soon become clear, an out-of-the-box root-finding approach can be unduly slow or even fraught with numerical peril, effects that become more pronounced with increasing data dimensionality. Let us, therefore, consider a slightly more general equation (we also drop the accent on κ):

$$\begin{aligned} &g(a, c; \kappa) := \frac{M'(a, c; \kappa)}{M(a, c; \kappa)} = r \\ &c > a > 0, \qquad 0 \leq r \leq 1. \end{aligned} \tag{2.7}$$

3. Solving for κ

In this section we present two different solutions to (2.7). The first is the "obvious" method based on a Newton–Raphson root-finder. The second method is the key numerical contribution of this paper: a method that computes a closed-form approximate solution to (2.7), thereby requiring merely a few floating-point operations!

3.1. Newton–Raphson

Although we establish this fact not until Section 3.2, suppose for the moment that (2.7) *does* have a solution. Further, assume that by bisection or otherwise, we have bracketed the root κ to be within an interval and are thus ready to invoke the Newton–Raphson method.

Starting at κ_0, Newton–Raphson solves the equation $g(a, c; \kappa) - r = 0$ by iterating

$$\kappa_{n+1} = \kappa_n - \frac{g(a, c; \kappa_n) - r}{g'(a, c; \kappa_n)}, \quad n = 0, 1, \ldots. \tag{3.1}$$

This iteration may be simplified by rewriting $g'(a, c; \kappa)$. First note that

$$g'(a, c; \kappa) = \frac{M''(a, c; \kappa)}{M(a, b; \kappa)} - \left(\frac{M'(a, c; \kappa)}{M(a, c; \kappa)}\right)^2, \tag{3.2}$$

then, recall the following two identities

$$M''(a, c; \kappa) = \frac{a(a+1)}{c(c+1)} M(a+2, c+2; \kappa); \tag{3.3}$$

$$M(a+2, c+2; \kappa) = \frac{(c+1)(-c+\kappa)}{(a+1)\kappa} M(a+1, c+1; \kappa) + \frac{(c+1)c}{(a+1)\kappa} M(a, c; \kappa). \tag{3.4}$$

Now, use both (3.3) and (3.4) to rewrite the derivative (3.2) as

$$g'(a, c; \kappa) = (1 - c/\kappa) g(a, c; \kappa) + (a/\kappa) - (g(a, c; \kappa))^2. \tag{3.5}$$

The main consequence of these simplifications is that iteration (3.1) can be implemented with only *one* evaluation of the ratio $g(a, c; \kappa_n) = M'(a, c; \kappa_n)/M(a, c; \kappa_n)$. Efficiently computing this ratio is a *non-trivial* task in itself; an insight into this difficulty is offered by observations in [9,10]. In the worst case, one may have to compute the numerator and denominator separately (using multi-precision floating point arithmetic), and then divide. Doing so can require several million extended precision floating point operations, which is very undesirable.

3.2. Closed-form approximation for (2.7)

We now derive two-sided bounds which will lead to a closed-form approximation to the solution of (2.7). This approximation, while marginally less accurate than the one via Newton–Raphson, should suffice for most uses. Moreover, it is incomparably faster to compute as it is in closed-form.

Before proceeding to the details, let us look at a little history. For 2–3 dimensional data, or under very restrictive assumptions on κ or r, some approximations had been previously obtained [16]. Due to their restrictive assumptions, these approximations have limited applicability, especially for high-dimensional data, where these assumptions are often violated [3]. Recently Bijral et al. [4] followed the technique of [3] to essentially obtain the *ad-hoc* approximation (actually particularly for the case $a = 1/2$)

$$BBG(r) := \frac{cr - a}{r(1-r)} + \frac{r}{2c(1-r)}, \tag{3.6}$$

which they observed to be quite accurate. However, (3.6) lacks theoretical justification; other approximations were presented in [18], though again only *ad-hoc*.

Below we present new approximations for κ that are theoretically well-motivated and also numerically more accurate. Key to obtaining these approximations are a set of bounds localising κ, and we present these in a series of theorems below. However, space restrictions have forced us to omit some proofs as they are quite technically involved. The included proofs have also been trimmed. The reader is referred to the accompanying arXiv report [19] for complete proofs.

3.2.1. Existence and uniqueness

The following theorem shows that the function $g(a, c; \kappa)$ is strictly increasing.

Theorem 3.1. *Let $c > a > 0$, and $\kappa \in \mathbb{R}$. Then the function $\kappa \to g(a, c; \kappa)$ is monotone increasing from $g(a, c; -\infty) = 0$ to $g(a, c; \infty) = 1$.*

Proof. Since $g(a, c; \kappa) = (a/c) f_1(\kappa)$, where f_μ is defined in (A.7), this theorem is a direct consequence of Theorem A.4. □

Hence the equation $g(a, c; \kappa) = r$ has a unique solution for each $0 < r < 1$. This solution is negative if $0 < r < a/c$ and positive if $a/c < r < 1$. Let us now localise this solution to a narrow interval by deriving tight bounds on it.

3.2.2. Bounds on the solution κ

Deriving tight bounds for κ is key to obtaining our new theoretically well-defined numerical approximations; moreover, these approximations are easy to compute because the bounds are given in closed form.

Theorem 3.2. *Let the solution to $g(a, c; \kappa) = r$ be denoted by $\kappa(r)$. Consider the following three bounds:*

$$\text{(lower bound) } L(r) = \frac{rc - a}{r(1-r)} \left(1 + \frac{1-r}{c-a}\right), \tag{3.7}$$

$$\text{(bound) } B(r) = \frac{rc - a}{2r(1-r)} \left(1 + \sqrt{1 + \frac{4(c+1)r(1-r)}{a(c-a)}}\right), \tag{3.8}$$

$$\text{(upper bound) } U(r) = \frac{rc - a}{r(1-r)} \left(1 + \frac{r}{a}\right). \tag{3.9}$$

Let $c > a > 0$, and $\kappa(r)$ be the solution (2.7). Then, we have

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

$$L(r) < \kappa(r) < B(r) < U(r), \tag{3.10}$$

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

$$L(r) < B(r) < \kappa(r) < U(r). \tag{3.11}$$

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 $r = 1$.

Proof. The proofs of parts 1 and 2 are given in Theorems A.5 and A.6 (see the Appendix), respectively. Part 3 is trivial. It is easy to see that $\lim_{r \to 0,1} U(r)/L(r) = 1$, so from inequalities (3.10) and (3.11), it follows that

$$\lim_{r \to 0,1} \frac{L(r)}{\kappa(r)} = \lim_{r \to 0,1} \frac{B(r)}{\kappa(r)} = \lim_{r \to 0,1} \frac{U(r)}{\kappa(r)} = 1. \quad \square$$

More precise asymptotic characterisations of the approximations L, B and U are given in Section 3.2.4.

Table 1
Summary of various approximations.

Point	Approx.			
	$L(r)$	$B(r)$	$U(r)$	$BBG(r)$
$r = 0$	Correct of order 1	Correct of order 2	Correct of order 3	Correct of order 2
$r = a/c$	Correct of order 2	Correct of order 3	Correct of order 2	Incorrect
$r = 1$	Correct of order 3	Correct of order 2	Correct of order 1	Correct of order 1

3.2.3. BBG approximation

Our bounds above also provide some insight into the previous approximation $BBG(r)$ of [4] given by (3.6). Specifically, we check whether $BBG(r)$ satisfies the lower and upper bounds from Theorem 3.2.

To see when $BBG(r)$ violates the lower bound, solve $L(r) > BBG(r)$ for r to obtain

$$\frac{2c^2 + a - \sqrt{(2c^2 - a)(2c^2 - a - 8ac)}}{2(2c^2 - a + c)} < r < \frac{2c^2 + a + \sqrt{(2c^2 - a)(2c^2 - a - 8ac)}}{2(2c^2 - a + c)}.$$

For the Watson case $a = 1/2$; this means that $BBG(r)$ violates the lower bound and *underestimates* the solution for $r \in (0.11, 0.81)$ if $c = 5$; for $r \in (0.0528, 0.904)$ if $c = 10$; for $r \in (0.00503, 0.99)$ if $c = 100$; for $r \in (0.00050025, 0.999)$ if $c = 1000$. This fact is also reflected in Fig. 2.

To see when $BBG(r)$ violates the upper bound, solve $BBG(r) > U(r)$ for r to obtain

$$r < \frac{2ac}{2c^2 - a}.$$

For the Watson case $a = 1/2$; this means that $BBG(r)$ violates the upper bound and *overestimates* the solution for $r \in (0, 0.1)$ if $c = 5$; for $r \in (0, 0.05)$ if $c = 10$; for $r \in (0, 0.005)$ if $c = 100$; for $r \in (0, 0.0005)$ if $c = 1000$.

What do these violations imply? They show that a combination of $L(r)$ and $U(r)$ is guaranteed to give a better approximation than $BBG(r)$ for nearly all $r \in (0, 1)$ except for a very small neighbourhood of the point where $BBG(r)$ intersects $\kappa(r)$.

3.2.4. Asymptotic precision of the approximations

Let us now look more precisely at how the various approximations behave at limiting values of r. There are three points where we can compute asymptotics: $r = 0$, $r = a/c$, and $r = 1$. First, we assess how $\kappa(r)$ itself behaves.

Theorem 3.3. *Let $c > a > 0$, $r \in (0, 1)$; let $\kappa(r)$ be the solution to $g(a, c; \kappa) = r$. Then,*

$$\kappa(r) = -\frac{a}{r} + (c - a - 1) + \frac{(c - a - 1)(1 + a)}{a} r + O(r^2), \quad r \to 0, \tag{3.12}$$

$$\kappa(r) = \left(r - \frac{a}{c}\right)\left\{\frac{c^2(1 + c)}{a(c - a)} + \frac{c^3(1 + c)^2(2a - c)}{a^2(c - a)^2(c + 2)}\left(r - \frac{a}{c}\right) + O\left(\left(r - \frac{a}{c}\right)^2\right)\right\}, \quad r \to \frac{a}{c} \tag{3.13}$$

$$\kappa(r) = \frac{c - a}{1 - r} + 1 - a + \frac{(a - 1)(a - c - 1)}{c - a}(1 - r) + O((1 - r)^2), \quad r \to 1. \tag{3.14}$$

Proof. The proof hinges on Lagrange inversion and its guises for unbounded values. Details can be found in [19]. □

We can compute asymptotic expansions for the various approximations by standard Laurent expansion. We summarise the results in Table 1 below; the formulae for the associated approximations are omitted due to space concerns, and may be found in [19].

Table 1 uses the following terminology: (i) we call an approximation $f(r)$ to be *incorrect* around $r = \alpha$, if $f(r)/\kappa(r) \to 0, \infty$ as $r \to \alpha$; (ii) we say $f(r)$ is *correct of order* 1 around $r = \alpha$, if $f(r)/\kappa(r) \to C$ such that $C \neq 0, \infty$ as $r \to \alpha$; (iii) we say $f(r)$ is *correct of order* 2 around $r = \alpha$ if $f(r)/\kappa(r) = 1 + O(r - \alpha)$ as $r \to \alpha$; and (iv) $f(r)$ is *correct of order* 3 around $r = \alpha$ if $f(r)/\kappa(r) = 1 + O((r - \alpha)^2)$ as $r \to \alpha$.

No matter how we count the total "order of correctness" it is clear from Table 1 that our approximations are superior to that of [4].

The table shows that actually $L(r)$ and $U(r)$ can be viewed as three-point 2/2] Padé approximations to $\kappa(r)$ at $r = 0$ and $r = a/c$ and $r = 1$ with different orders at different points, while $B(r)$ is a special non-rational three point approximation with even higher total order of contact.

Moreover, since we not only give the order of correctness but also prove the inequalities, we always know exactly which approximation underestimates $\kappa(r)$ and which overestimates $\kappa(r)$. Such information might be important to some applications. The approximation of [4] is less precise and does not satisfy such inequalities. Also, note that all the above facts are equally true in the Watson case $a = 1/2$.

4. Application to mixture modelling and clustering

Now that we have shown how to compute maximum-likelihood parameter estimates, we proceed onto *mixture-modelling* for mW distributions.

Suppose we observe the set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{P}^{p-1}\}$ of i.i.d. samples. We wish to model this set using a mixture of K mW distributions. Let $W_p(\boldsymbol{x}|\boldsymbol{\mu}_j, \kappa_j)$ be the density of the j-th mixture component, and π_j its prior $(1 \leq j \leq K)$ — then, for observation $\boldsymbol{x}_i$ we have the density

$$f(\boldsymbol{x}_i|\boldsymbol{\mu}_1, \kappa_1, \ldots, \boldsymbol{\mu}_K, \kappa_K) = \sum_{j=1}^{K} \pi_j W_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \kappa_j).$$

The corresponding log-likelihood for the entire dataset $\mathcal{X}$ is given by

$$\mathcal{L}(\mathcal{X}; \boldsymbol{\mu}_1, \kappa_1, \ldots, \boldsymbol{\mu}_K, \kappa_K) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{K} \pi_j W_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \kappa_j) \right). \tag{4.1}$$

To maximise the log-likelihood, we follow a standard Expectation Maximisation (EM) procedure [6]. To that end, first bound $\mathcal{L}$ from below as

$$\mathcal{L}(\mathcal{X}; \boldsymbol{\mu}_1, \kappa_1, \ldots, \boldsymbol{\mu}_K, \kappa_K) \geq \sum_{ij} \beta_{ij} \ln \frac{\pi_j W_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \kappa_j)}{\beta_{ij}}, \tag{4.2}$$

where β_{ij} is the *posterior* probability (for $\boldsymbol{x}_i$, given component j), and it is defined by the *E-Step*:

$$\beta_{ij} = \frac{\pi_j W_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \kappa_j)}{\sum_l \pi_l W_p(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \kappa_l)}. \tag{4.3}$$

Maximising the lower-bound (4.2) subject to $\boldsymbol{\mu}_j^\top \boldsymbol{\mu}_j = 1$, yields the *M-Step*:

$$\boldsymbol{\mu}_j = \boldsymbol{s}_1^j \quad \text{if } \kappa_j > 0, \qquad \boldsymbol{\mu}_j = \boldsymbol{s}_p^j \quad \text{if } \kappa_j < 0, \tag{4.4}$$

$$\kappa_j = g^{-1}(1/2, p/2, r_j), \quad \text{where } r_j = \boldsymbol{\mu}_j^\top \boldsymbol{S}^j \boldsymbol{\mu}_j, \tag{4.5}$$

$$\pi_j = \frac{1}{n} \sum_i \beta_{ij},$$

where $\boldsymbol{s}_i^j$ denotes the eigenvector corresponding to eigenvalue λ_i (where $\lambda_1 \geq \cdots \geq \lambda_p$) of the *weighted-scatter matrix:*

$$\boldsymbol{S}^j = \frac{1}{\sum_i \beta_{ij}} \sum_i \beta_{ij} \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

Now we can iterate between (4.3)–(4.5) to obtain an EM algorithm. Pseudo-code for such a procedure is shown below as Algorithm 1.

Note: Hard assignments. We note that as usual, to reduce the computational burden, we can replace the E-step (4.3) by the standard *hard-assignment* heuristic:

$$\beta_{ij} = \begin{cases} 1, & \text{if } j = \operatorname{argmax}_{j'} \ln \pi_{j'} + \ln W_p(\boldsymbol{x}_i|\boldsymbol{\mu}_{j'}, \kappa_{j'}), \\ 0, & \text{otherwise.} \end{cases} \tag{4.6}$$

The corresponding M-Step also simplifies considerably. Such hard-assignments maximise a lower-bound on the incomplete log-likelihood, and yield *partitional-clustering* algorithms (in fact, we show experimental results in Section 5.2 where we cluster data using a partitional-clustering algorithm based on this hard-assignment heuristic).

4.1. Diametrical clustering

We now turn to the diametrical clustering algorithm of [7], and show that it is merely a special case of the mixture-model described above. Diametrical clustering is motivated by the need to group together correlated and anti-correlated data points (see Fig. 1 for an illustration). For data normalised to have unit euclidean norm, such clustering treats *diametrically* opposite points equivalently. In other words, $\boldsymbol{x}$ lies on the projective plane. Therefore, a natural question is whether diametrical clustering is related to Watson distributions, and if so, how?

The answer to this question will become apparent once we recall the diametrical clustering algorithm (shown as Algorithm 2) of [7]. In Algorithm 2 we have labelled the "E-Step" and the "M-Step". These two steps are simplified instances

Input: $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n : \text{ where each } \boldsymbol{x}_i \in \mathbb{P}^{p-1}\}$, K: number of components
Output: Parameter estimates π_j, $\boldsymbol{\mu}_j$, and κ_j, for $1 \leq j \leq K$
Initialise π_j, $\boldsymbol{\mu}_j$, κ_j for $1 \leq j \leq K$
while *not converged* **do**
  {Perform the E-step of EM}
  foreach *i and j* **do**
    Compute β_{ij} using (4.3) (or via (4.6) if using hard-assignments)
  end
  {Perform the M-step of EM}
  for $j = 1$ *to* K **do**
    $\pi_j \leftarrow \frac{1}{n}\sum_{i=1}^{n} \beta_{ij}$
    Compute $\boldsymbol{\mu}_j$ using (4.4)
    Compute κ_j using (4.5)
  end
end

Algorithm 1: EM Algorithm for mixture of Watson (moW)

Input: $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n : \boldsymbol{x}_i \in \mathbb{P}^{p-1}\}$, K: number of clusters
Output: A partition $\{\mathcal{X}_j : 1 \leq j \leq K\}$ of $\mathcal{X}$, and centroids $\boldsymbol{\mu}_j$
Initialise $\boldsymbol{\mu}_j$ for $1 \leq j \leq K$
while not converged **do**
  E-step:
  Set $\mathcal{X}_j \leftarrow \emptyset$ for $1 \leq j \leq K$
  for $i = 1$ *to* n **do**
    $\mathcal{X}_j \leftarrow \mathcal{X}_j \cup \{\boldsymbol{x}_i\}$ where $j = \operatorname{argmax}_{1 \leq h \leq K} (\boldsymbol{x}_i^\top \boldsymbol{\mu}_h)^2$
  end
  M-step:
  for $j = 1$ *to* K **do**
    $\boldsymbol{A}_j = \sum_{\boldsymbol{x}_i \in \mathcal{X}_j} \boldsymbol{x}_i \boldsymbol{x}_i^\top$
    $\boldsymbol{\mu}_j \leftarrow \boldsymbol{A}_j \boldsymbol{\mu}_j / \|\boldsymbol{A}_j \boldsymbol{\mu}_j\|$
  end
end

Algorithm 2: Diametrical Clustering

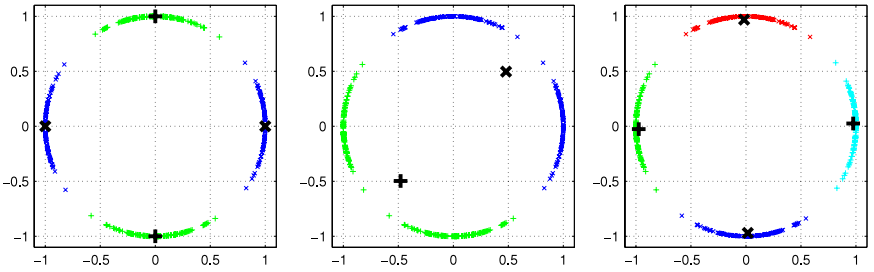

Fig. 1. The left panel shows axially symmetric data that has two clusters (centroids are indicated by '+' and '×'). The middle and right panels show clustering yielded by (Euclidean) K-means (note that the centroids fail to lie on the circle in this case) with $K = 2$ and $K = 4$, respectively. Diametrical clustering recovers the true clusters in the left panel.

of the E-step (4.3) (alternatively (4.6)) and M-step (4.4). To see why, consider the E-step (4.3). If $\kappa_j \to \infty$, then 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 $(\boldsymbol{\mu}_j^\top \boldsymbol{x}_i)^2$ is maximised — this is precisely the choice used in the E-step of Algorithm 2. With binary values for β_{ij}, the M-Step (4.4) also reduces to the version followed by Algorithm 2.

An alternative, perhaps better view is obtained by regarding diametrical clustering as a special case of mixture-modelling where a hard-assignment rule is used. Now, if all mixture components have the *same, positive* concentration parameter κ, then while computing β_{ij} via (4.6) we may ignore κ altogether, which reduces Algorithm 1 to Algorithm 2.

Given this interpretation of diametrical clustering, it is natural to expect that the additional modelling power offered by mixtures of Watson distributions might lead to better clustering. This is indeed the case, as indicated by some of our experiments in Section 5.2 below, where we show that merely including the concentration parameter κ can lead to improved clustering accuracies, or to clusters with higher quality (in a sense that will be made more precise below).

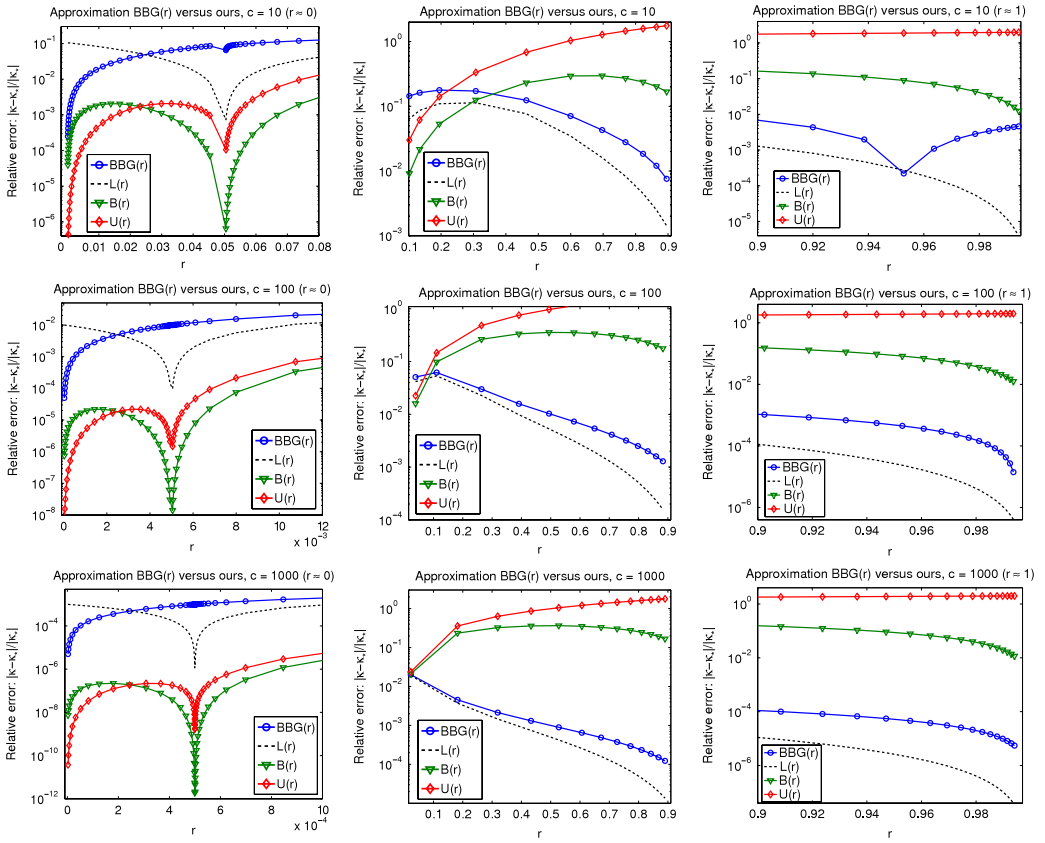

Fig. 2. Relative errors $|\hat{\kappa} - \kappa_*|/|\kappa_*|$ of $BBG(r)$, $L(r)$, $B(r)$, and $U(r)$ for $c \in \{10, 100, 1000\}$ as r varies between $(0, 1)$. The left column shows errors for "small" r (i.e., r close to 0), the middle column shows errors for "mid-range" r, and the last column shows errors for the "high" range ($r \approx 1$).

5. Experiments

We now come to numerical results to assess the methods presented. We divide our experiments into two groups. The first group comprises numerical results that illustrate accuracy of our approximation to κ. The second group supports our claim that the extra modelling power offered by moWs also translates into better clustering results.

5.1. Estimating κ

We show two representative experiments to illustrate the accuracy of our approximations. The first set (Section 5.1.1) compares our approximation with that of [4], as given by (3.6). This set considers the Watson case, namely $a = 1/2$ and varying dimensionality $c = p/2$. The second set (Section 5.1.2) of experiments shows a sampling of results for a few values of c and κ as the parameter a is varied. This set illustrates how well our approximations behave for the general nonlinear equation (2.7).

5.1.1. Comparison with the BBG approximation for the Watson case

Here we fix $a = 1/2$, and vary c on an exponentially spaced grid ranging from $c = 10$ to $c = 10^4$. For each value of c, we generate geometrically spaced values of the "true" κ_* in the range $[-200c, 200c]$. For each choice of κ_* picked within this range, we compute the ratio $r = g(1/2, c, \kappa_*)$ (using MATHEMATICA for high precision). Then, given $a = 1/2$, c, and r, we estimate κ_* by solving $\kappa \approx g^{-1}(1/2, c, r)$ using $BBG(r)$, $L(r)$, $B(r)$, and $U(r)$, given by (3.6), (3.7), (3.8), and (3.9), respectively.

Fig. 2 shows the results of computing these approximations. From the plots we see that although approximation $BBG(r)$ is quite good and more accurate than $B(r)$ and $U(r)$ for $r \approx 1$, approximation $L(r)$ is more accurate across almost the whole range of dimensions and r values. In contrast, for small r, $BBG(r)$ can be more accurate than $L(r)$, but in this case both $U(r)$ and $B(r)$ are much more accurate.

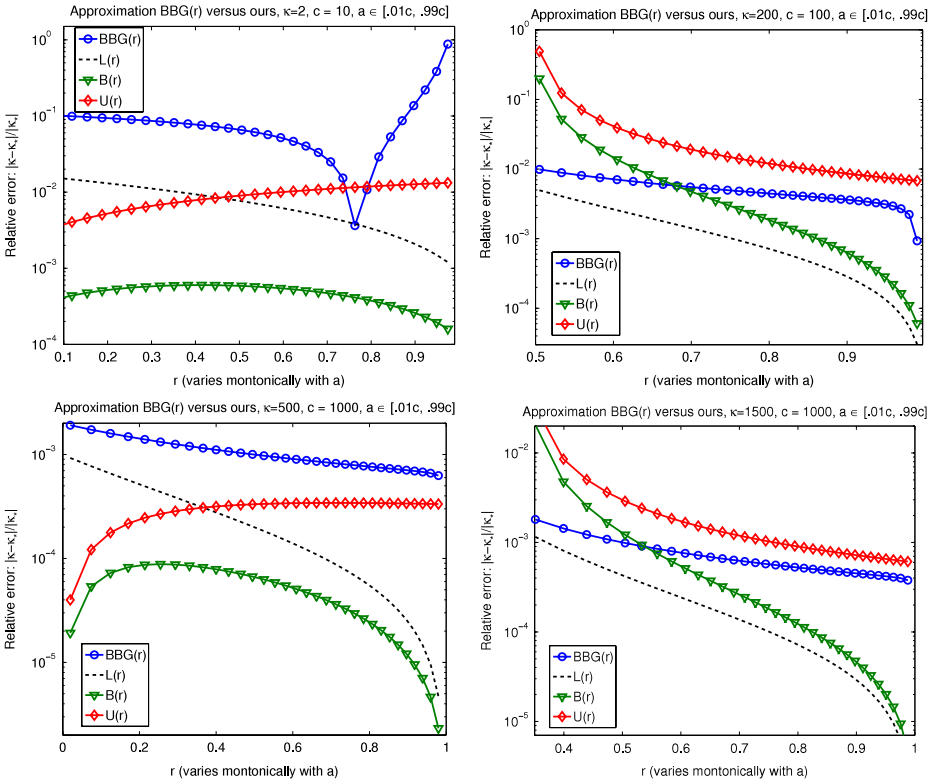

Fig. 3. Relative errors of $BBG(r)$, $L(r)$, $B(r)$, and $U(r)$ for different sets of c and κ values, as a is varied from $0.01c$ to $0.99c$.

5.1.2. Comparisons of the approximation for fixed c and varying a

In our next set of experiments, we chose a few values of c and κ (see Fig. 3), and varied a linearly to lie in the range $[0.01c, 0.99c]$. Fig. 3 reports the relative errors of approximation incurred by the various approximations.

From the plots it is clear that one of $L(r)$, $B(r)$, or $U(r)$ always yields results more accurate than $BBG(r)$. The various results suggest the following rough rule-of-thumb: prefer $U(r)$ for $0 < r < a/(2c)$, prefer $B(r)$ for $a/(2c) \leq r < 2a/\sqrt{c}$ and prefer $L(r)$ for $2a/\sqrt{c} \leq r < 1$.

5.2. Clustering using mW distributions

Now we turn to our second set of experiments. Below we show results of two experiments: (i) with synthetic data, where a desired "true-clustering" is known; and (ii) with gene expression data for which previously axially symmetric clusters have been considered.

For both our experiments, we compare moW (Algorithm 1 with (4.6) for the E-step) against the diametrical clustering procedure of [7]. The key aim of the experiments is to show that the extra modelling power offered by a mixture of mW distributions can provide clustering results better than plain diametrical clustering.

5.2.1. Synthetic data

We generated data that merely exhibit axial symmetry and have varying degrees of concentration around given mean directions. Since both the diametrical method as well as moW model axial symmetry they can be fairly compared on this data. The distinction comes, however, where moW further models concentration (via κ), and in case the generated data is sufficiently concentrated, this modelling translates into empirically superior performance. Naturally, to avoid unfairly skewing results in favour of moW, we do not compare it against diametrical clustering on synthetic data sampled from a mixture of mW distributions as moW explicitly optimises such a model.

For our data generation we need to sample points from $W_p(\kappa, \boldsymbol{\mu})$, for which we invoke a simplified version of the powerful Gibbs sampler of [11] that can simulate Bingham–von Mises–Fisher distributions. We note here that Bingham distribution is parametrised by a matrix $\boldsymbol{A}$, and to use it for sampling Watson distributions, we merely need to realise that $\boldsymbol{A} = \kappa\boldsymbol{\mu}\boldsymbol{\mu}^\top$.

With the sampling code in hand, we generate synthetic datasets with varying concentration as follows. First, two random unit vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{P}^{29}$ are selected. Then, we fix $\kappa_1 = 3$ and sample 200 points from $W_3(\kappa_1, \boldsymbol{\mu}_1)$. Next, we vary κ_2 in the

Table 2
Percentages of accurately clustered points for diametrical clustering vs. moW (over 10 runs). Since this is simulated data, we knew the cluster labels. The accuracy is then computed by matching the predicted labels with the known ones. In line with the theory, with increasing concentration the modelling power offered by moW shows a clear advantage over ordinary diametrical clustering.

κ_2	Diametrical (avg/best/worst)—%	moW (avg/best/worst)—%
3	52.65 / 56.50 / 51.50	51.65 / 53.50 / 50.50
10	52.75 / 56.00 / 50.50	54.10 / 57.00 / 50.00
20	57.60 / 64.00 / 51.50	74.45 / 87.00 / 63.50
50	66.00 / 78.50 / 50.00	99.50 / 99.50 / 99.50
100	71.20 / 81.00 / 55.00	100.00 / 100.00 / 100.00

set {3, 10, 20, 50, 100}, and generate 200 points for each value of κ_2 by sampling from $W_p(\kappa_2, \boldsymbol{\mu}_2)$. Finally, by mixing the κ_1 component with each of the five κ_2 components we obtain five datasets $\mathcal{X}_t$ $(1 \leq t \leq 5)$.

Each of these five datasets is then clustered into two clusters, using moW and diametrical clustering. Both algorithms are run ten times each to smooth out the effect of random initialisations. Table 2 shows the results of clustering by displaying the accuracy which measures the percentage of data points that were assigned to the "true" clusters (i.e., the true components in the mixture). The accuracies strongly indicate that explicit modelling of concentration leads to better clustering as κ_2 increases. In other words, larger κ_2 makes points from the second cluster more concentrated around $\pm\boldsymbol{\mu}_2$, thereby allowing easier separation between the clusters.

5.2.2. Real data

We now compare clustering results of moW with those of diametrical clustering on three gene microarray datasets that were also used in the original diametrical clustering paper [7]. These datasets are: (i) Human Fibroblasts [13]; (ii) Yeast Cell Cycle [17]; and (iii) Rosetta yeast [12]. The respective matrix sizes that we used were: (i) 517×12; (ii) 696×82; and (iii) 900×300 (these 900 genes were randomly selected from the original 5245).

Since we do not have ground-truth clusterings for these datasets, we validate our results using internal measures. Specifically, we compute two scores: *homogeneity* and *separation*, which are defined below by H_{avg} and S_{avg}, respectively. Let $\mathcal{X}_j \subset \mathcal{X}$ denote cluster j; then we define

$$H_{\text{avg}} = \frac{1}{n}\sum_{j=1}^{K}\sum_{\boldsymbol{x}_i \in \mathcal{X}_j} (\boldsymbol{x}_i^\top \boldsymbol{\mu}_j)^2, \tag{5.1}$$

$$S_{\text{avg}} = \frac{1}{\sum_{j\neq l} |\mathcal{X}_j||\mathcal{X}_l|}\sum_{j\neq l} |\mathcal{X}_j||\mathcal{X}_l| \min(\boldsymbol{\mu}_j^\top \boldsymbol{\mu}_l, -\boldsymbol{\mu}_j^\top \boldsymbol{\mu}_l). \tag{5.2}$$

We note a slight departure from the standard in our definitions above. In (5.1), instead of summing over $\boldsymbol{x}_i^\top \boldsymbol{\mu}_j$, we sum over their squares, while in (5.2), instead of $\boldsymbol{\mu}_j^\top \boldsymbol{\mu}_l$, we use $\min(\boldsymbol{\mu}_j^\top \boldsymbol{\mu}_l, -\boldsymbol{\mu}_j^\top \boldsymbol{\mu}_l)$ because for us $+\boldsymbol{\mu}_j$ and $-\boldsymbol{\mu}_j$ represent the same cluster.

We note that diametrical clustering optimises precisely the criterion (5.1), and is thus favoured by our criterion. Higher values of H_{avg} mean that the clusters have higher intra-cluster cohesiveness, and thus are "better" clusters. In contrast, lower values of S_{avg} mean that the inter-cluster dissimilarity is high, i.e., better separated clusters.

Table 3 shows results yielded by diametrical clustering and moW on the three different gene datasets. For each dataset, we show results for two values of K. The H_{avg} values indicate that moW yields clusters having approximately the same intra-cluster cohesiveness as diametrical. However, moW attains better inter-cluster separation as it more frequently leads to lower S_{avg} values.

6. Conclusions

We studied the multivariate Watson distribution, a fundamental tool for modelling axially symmetric data. We solved the difficult nonlinear equations that arise in maximum-likelihood parameter estimation. In high-dimensions these equations pose severe numerical challenges. We derived tight two-sided bounds that led to approximate solutions to these equations; we also showed our solutions to be accurate. We applied our results to mixture-modelling with Watson distributions and consequently uncovered a connection to the diametrical clustering algorithm of [7]. Our experiments showed that for clustering axially symmetric data, the additional modelling power offered by mixtures of Watson distributions can lead to better clustering. Further refinements to the clustering procedure, as well as other applications of Watson mixtures in high-dimensional settings is left as a task for the future.

Acknowledgments

The first author thanks Prateek Jain for initial discussions related to Watson distributions. The second author acknowledges the support of the Russian Basic Research Fund (grant 11-01-00038-a).

Table 3
Clustering accuracy on gene-expression datasets (over 10 runs). Noticeable differences (i.e., >0.02) between the algorithms are highlighted in bold.

Method	Diametrical (avg/best/worst)	moW (avg/best/worst)
Yeast-4		
Homogeneity	0.38 / 0.38 / 0.38	0.37 / 0.37 / 0.37
Separation	−0.00 / −0.23 / 0.24	**−0.04** / −0.23 / **0.20**
Yeast-6		
Homogeneity	0.41 / 0.41 / 0.40	0.41 / 0.41 / 0.40
Separation	−0.06 / −0.15 / 0.14	−0.07 /**−0.20** / 0.13
Rosetta-2		
Homogeneity	0.16 / 0.17 / 0.16	0.16 / 0.17 / 0.16
Separation	0.24 / 0.08 / 0.28	**−0.20** / **−0.28** / **0.09**
Rosetta-4		
Homogeneity	0.23 / 0.23 / 0.23	0.23 / 0.23 / 0.23
Separation	−0.01 / −0.08 / 0.16	−0.03 / −0.09 / **0.12**
Fibroblast-2		
Homogeneity	0.70 / 0.70 / 0.70	0.70 / 0.70 / 0.70
Separation	0.26 / −0.65 / 0.65	**−0.01** / −0.65 / 0.65
Fibroblast-5		
Homogeneity	0.78 / 0.78 / **0.78**	0.76 / 0.76 / 0.75
Separation	−0.05 / −0.28 / 0.40	**−0.12** / −0.30 / **0.35**

Appendix. Mathematical details

This appendix presents a few of the most relevant technical details omitted from the main text. Due to space limitations we cannot present all the proofs and appurtenant details; these may be found in the longer version of the paper [19].

We list below some identities for M that we will need for our analysis. To ease the notational burden, we also use the shorthand $M_i \equiv M(a+i, c+i, x)$.

$$\frac{d^n}{dx^n}M_0 = \frac{a^{\overline{n}}}{c^{\overline{n}}}M_n, \tag{A.1}$$

$$M_1 = \frac{c(1-c+x)}{ax}M_0 + \frac{c(c-1)}{ax}M_{-1}, \tag{A.2}$$

$$(c-a)xM(a+2, c+3, x) = (c+1)(c+2)[M_2 - M_1], \tag{A.3}$$

$$xM(a+2, c+3, x) = (c+2)[M_2 - M(a+1, c+2, x)]. \tag{A.4}$$

Identity (A.1) follows inductively; (A.2) is from [5, 16.1.9c]; (A.3) is obtained by combining [8, formula 6.4(5)] with [8, formula 6.4(4)]; and (A.4) is from [8, formula 6.4(5)]. The following lemma has been derived by the second author in [14, Lemma 1].

Lemma A.1. *The Kummer function satisfies the identity*

$$M_1^2 - M_2M_0 = \frac{(c-a)x}{c+1}\left[\frac{1}{c+1}M(a+1, c+2, x)^2 - \frac{1}{c+2}M(a+2, c+3, x)M(a, c+1, x) + \frac{1}{c(c+1)}M(a+1, c+2, x)M(a+2, c+2, x)\right]. \tag{A.5}$$

The central object of study in this paper is the *Kummer-ratio*:

$$g(x) = g(a, c; x) := \frac{M'(a, c, x)}{M(a, c, x)} = \frac{a}{c}\frac{M(a+1, c+1, x)}{M(a, c, x)}. \tag{A.6}$$

In the sequel, it will be useful to use the slightly more general function

$$f_\mu(x) := \frac{M(a+\mu, c+\mu, x)}{M(a, c, x)}, \quad \mu > 0, \tag{A.7}$$

so that $g(x) = (a/c)f_1(x)$.

Lemma A.2 (*Log-convexity*)**.** *Let $c > a > 0$ and $x \geq 0$. Then the function*

$$\mu \mapsto \frac{\Gamma(a+\mu)}{\Gamma(c+\mu)}M(a+\mu, c+\mu, x) = \sum_{k=0}^{\infty}\frac{\Gamma(a+\mu+k)}{\Gamma(c+\mu+k)}\frac{x^k}{k!} =: h_{a,c}(\mu; x)$$

is strictly log-convex on $[0, \infty)$ (note that h is a function of μ).

Proof. Write the power-series expansion in x for $h_{a,c}(\mu; x)$ as

$$h_{a,c}(\mu; x) = \sum_{k=0}^{\infty} h_k(a, c, \mu)\frac{x^k}{k!}, \qquad h_k(a, c, \mu) = \frac{\Gamma(a+\mu+k)}{\Gamma(c+\mu+k)}.$$

Since log-convexity is additive, it is sufficient to prove that $\mu \mapsto h_k(a, c, \mu)$ is log-convex. For this we compute the second-derivative

$$\frac{\partial^2}{\partial\mu^2} \ln h_k(a, c, \mu) = \psi'(a+\mu+k) - \psi'(c+\mu+k),$$

where ψ is the logarithmic derivative of the gamma function. We need to show that this expression is positive when $c > a > 0$, $k \geq 0$ and $\mu \geq 0$. Differentiating the Gauss formula [1, Theorem 1.6.1] twice we get

$$\psi''(x) = -\int_0^{\infty} \frac{t^2 e^{-tx}}{1-e^{-t}} dt < 0.$$

Hence the function $\psi'(x)$ is decreasing and our claim follows. □

Lemma A.3. *Let $c > a > 0$, and $x \geq 0$. Then the function*

$$\mu \mapsto \frac{\Gamma(a+\mu)}{\Gamma(c+\mu)} M(c-a, c+\mu, x) =: \hat{h}_{a,c}(\mu; x)$$

is strictly log-convex on $[0, \infty)$.

Proof. Using precisely the same argument as in the proof of Lemma A.2 we see that $\mu \mapsto \Gamma(a+\mu)/\Gamma(c+\mu)$ is log-convex. Next, the log-convexity of $\mu \mapsto M(c-a; c+\mu; x)$ has been proved by several authors (see, for instance, [15] and references therein). Thus multiplicativity of log-convexity completes the proof. □

With the last two lemmas in hand we are ready to prove the following theorem.

Theorem A.4 (*Monotonicity*)**.** *Let $c > a > 0$. Then, the function $x \mapsto f_\mu(x)$ is monotone increasing on $(-\infty, \infty)$, with $f_\mu(-\infty) = 0$ and $f_\mu(\infty) = \Gamma(c+\mu)\Gamma(a)/\big(\Gamma(c)\Gamma(a+\mu)\big)$.*

Proof. We divide the proof into two cases: (i) $x \geq 0$, and (ii) $x < 0$.

Case (i). It follows from (A.1) that

$$M_0^2 f'_\mu(x) = \frac{a+\mu}{c+\mu} M_{\mu+1} M_0 - \frac{a}{c} M_\mu M_1.$$

We need to show that the above expression is positive, which amounts to showing

$$\frac{a+\mu}{c+\mu}\frac{M_{\mu+1}}{M_\mu} > \frac{a}{c}\frac{M_1}{M_0} \Leftrightarrow \frac{[\Gamma(a+\mu+1)/\Gamma(c+\mu+1)]M_{\mu+1}}{[\Gamma(a+\mu)/\Gamma(c+\mu)]M_\mu} > \frac{[\Gamma(a+1)/\Gamma(c+1)]M_1}{[\Gamma(a)/\Gamma(c)]M_0}. \tag{A.8}$$

The last inequality follows from Lemma A.2. To see how, recall that if $\mu \mapsto h(\mu)$ is log-convex, then the function $\mu \mapsto h(\mu+\delta)/h(\mu)$ is increasing for each fixed $\delta > 0$, a fact that is easily verified by noting that when h is log-convex, its logarithmic derivative $h'(\mu)/h(\mu)$ is increasing, which immediately implies that the derivative of $h(\mu+\delta)/h(\mu)$ is positive. Thus, in particular applying this property to $h_{a,c}(\mu; x)$ with $\delta = 1$ we have

$$\frac{h_{a,c}(\mu+1; x)}{h_{a,c}(\mu; x)} > \frac{h_{a,c}(1; x)}{h_{a,c}(0; x)},$$

which is precisely the required inequality. This establishes the monotonicity. The value of $f_\mu(\infty)$ follows from the asymptotic formula given in [1, Corollary 4.2.3].

Case (ii). Let $x < 0$. Like in Case (i) we need to show that

$$\frac{[\Gamma(a+\mu+1)/\Gamma(c+\mu+1)]M_{\mu+1}}{[\Gamma(a+\mu)/\Gamma(c+\mu)]M_\mu} > \frac{[\Gamma(a+1)/\Gamma(c+1)]M_1}{[\Gamma(a)/\Gamma(c)]M_0} \tag{A.9}$$

but this time for $x < 0$. Apply the Kummer transformation $M(a; c; x) = e^x M(c-a; c; -x)$ and write $y = -x > 0$ to get

$$\frac{[\Gamma(a+\mu+1)/\Gamma(c+\mu+1)]M(c-a; c+\mu+1; y)}{[\Gamma(a+\mu)/\Gamma(c+\mu)]M(c-a; c+\mu; y)} > \frac{[\Gamma(a+1)/\Gamma(c+1)]M(c-a; c+1; y)}{[\Gamma(a)/\Gamma(c)]M(c-a; c; y)}.$$

Using the notation introduced in Lemma A.3 the last inequality becomes

$$\frac{\hat{h}_{a,c}(\mu+1;x)}{\hat{h}_{a,c}(\mu;x)} > \frac{\hat{h}_{a,c}(1;x)}{\hat{h}_{a,c}(0;x)},$$

which holds as a consequence of the log-convexity of $\mu \mapsto \hat{h}_{a,c}(\mu;x)$. Finally, combining the Kummer transformation with [1, Corollary 4.2.3] we get the value of $f_\mu(-\infty)$. □

Let us remind the reader that the functions $L(r)$, $B(r)$ and $U(r)$ are defined in (3.7), (3.8), (3.9), respectively.

Theorem A.5 (*Positive κ*). *Let $\kappa(r)$ be the solution to (2.7); $c > a > 0$, and $r \in (a/c, 1)$. Then, we have the bounds*

$$L(r) < \kappa(r) < B(r) < U(r) \tag{A.10}$$

Proof. *Lower-bound.* To simplify notation we use $x = \kappa(r)$ below. Set $r_1 = g(a+1, c+1; x)$; then replace $a \leftarrow a+1$, $c \leftarrow c+1$ and divide by M_1 in identity (A.2) to obtain

$$x = \frac{cr - a}{r(1 - r_1)}, \tag{A.11}$$

where as before $r = g(a, c, x)$. The lower bound in (A.10) is equivalent to

$$\frac{cr-a}{r(1-r_1)} > \frac{cr-a}{r(1-r)} + \frac{cr-a}{r(c-a)} \quad \text{or} \quad \frac{(c-a-1)r+1}{c-a+1-r} < r_1,$$

once we account for $cr - a > 0$ by our hypothesis. Plugging in the definitions of r and r_1 we get:

$$\frac{(c-a-1)aM_1 + cM_0}{(c-a+1)cM_0 - aM_1} < \frac{(a+1)M_2}{(c+1)M_1},$$

where as before we use $M_i = M(a+i, c+i, x)$. Cross-multiplying, we obtain

$$h(x) := c(c-a+1)(a+1)M_2M_0 - (c+1)(c-a-1)aM_1^2 - c(c+1)M_1M_0 - a(a+1)M_2M_1 > 0.$$

Now on noticing that $c(c-a+1)(a+1) = ac(c-a) + c(c+1)$ and $(c-a-1)(c+1)a = ac(c-a) - a(a+1)$, we can regroup $h(x)$ to get

$$h(x) = ac(c-a)[M_2M_0 - M_1^2] + (M_2 - M_1)[c(c+1)M_0 - a(a+1)M_1].$$

Next, using identity (A.3) for $M_2 - M_1$, formula (A.5) for $M_2M_0 - M_1^2$, the easily verifiable identity

$$c(c+1)M_0 - a(a+1)M_1 = \frac{ax(c-a)}{c+1}M(a+1, c+2, x) + (c-a)(c+a+1)M(a, c+1, x)$$

and the contiguous relation (A.4), $h(x)$ can be brought into the form

$$\frac{(c+1)h(x)}{(c-a)^2x} = \frac{(a+1)(c+1)}{c+2}M(a, c+1, x)M(a+2, c+3, x) - aM(a+1, c+2, x)^2.$$

Therefore, the condition $h(x) > 0$ is equivalent to (upon using the notation $c' = c+1$)

$$\frac{(a+1)}{(c'+1)}\frac{M(a+2, c'+2, x)}{M(a+1, c'+1, x)} > \frac{a}{c'}\frac{M(a+1, c'+1, x)}{M(a, c', x)}.$$

But this final inequality follows from Theorem A.4 by using $\mu = 1$ in (A.8).

Upper-bound for $\kappa(r)$. The lower-bound in (A.10) can be then rewritten as

$$\frac{cr-a}{r(1-r)}\left(1 + \frac{1-r}{c-a}\right) < x \Leftrightarrow (1-r)^2(x - c/(c-a)) - (1-r)(x+c-1) + c - a < 0.$$

The last inequality can be shown to imply

$$\frac{(c-a)(x+c-1) - (c-a)\sqrt{(1-x+c)^2 + 4ax}}{2((c-a)x - c)} < 1 - r.$$

Changing $a \to a+1$ and $c \to c+1$ here (recall that $b = c - a$) we get

$$0 < \frac{b(x+c) - b\sqrt{(x+c)^2 - 4(bx - c - 1)}}{2(bx - c - 1)} < 1 - r_1, \tag{A.12}$$

where as before $r_1 = g(a+1, c+1, x)$. The expression under square root is positive on inspection for both $bx - c - 1 > 0$ and $bx - c - 1 < 0$. Next, after suitably rewriting (A.11), we have

$$x = \frac{b - cq}{(1-q)(1-r_1)}.$$

Applying inequality (A.12) here, we obtain

$$x < \frac{2(b-cq)(bx-c-1)}{(1-q)b\left(x + c - \sqrt{(x+c)^2 - 4(bx-c-1)}\right)}.$$

Carefully solving this inequality for x we get the upper bound in (A.10).

The rightmost inequality. Verifying the inequality $B(r) < U(r)$ for $a/c < r < 1$ is a straightforward exercise. □

Theorem A.6 (*Negative κ*). *Let $\kappa(r)$ be the solution to (2.7), $c > a > 0$, and $r \in (0, a/c)$. Then, we have the following bounds:*

$$L(r) < B(r) < \kappa(r) < U(r) \tag{A.13}$$

Proof. The proof goes along the lines similar to those in the proof of Theorem A.5 but with many differences in technical details. For complete version see [19]. □

References

[1] G.E. Andrews, R. Askey, R. Roy, Special Functions, Cambridge University Press, 1999.
[2] A. Banerjee, I.S. Dhillon, J. Ghosh, S. Sra, Generative model-based clustering of directional data, in: Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-2003. 2003, pp. 19–28.
[3] A. Banerjee, I.S. Dhillon, J. Ghosh, S. Sra, Clustering on the unit hypersphere using von Mises–Fisher distributions, Journal of Machine Learning Research 6 (2005) 1345–1382.
[4] A. Bijral, M. Breitenbach, G.Z. Grudic, Mixture of watson distributions: a generative model for hyperspherical embeddings, in: Artificial Intelligence and Statistics, AISTATS 2007, 2007, pp. 35–42.
[5] A. Cuyt, V.B. Petersen, B. Verdonk, H. Waadeland, W.B. Jones, Handbook of Continued Fractions for Special Functions, Springer, 2008.
[6] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society 39 (1977).
[7] I.S. Dhillon, E.M. Marcotte, U. Roshan, Diametrical clustering for identifying anti-correlated gene clusters, Bioinformatics 19 (13) (2003) 1612–1619.
[8] A. Erdélyi, W. Magnus, F. Oberhettinger, F.G. Tricomi, Higher transcendental functions. Vol. 1, McGraw Hill, 1953.
[9] W. Gautschi, Anomalous convergence of a continued fraction for ratios of kummer functions, Mathematics of Computation 31 (140) (1977) 994–999.
[10] A. Gil, J. Segura, N.M. Temme, Numerical Methods for Special Functions, Cambridge University Press, 2007.
[11] P.D. Hoff, Simulation of the Matrix Bingham–von Mises–Fisher Distribution, with applications to multivariate and relational data, Journal of Computational and Graphical Statistics 18 (2) (2009) 438–456.
[12] T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, S.H. Friend, Functional discovery via a compendium of expression profiles, Cell 102 (2000) 109–126.
[13] V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C.F. Lee, J.M. Trent, L.M. Staudt, J. Hudson, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, P.O. Brown, The transcriptional program in the response of human fibroblasts to serum, Science 283 (5398) (1999) 83–87.
[14] D. Karp, Turán's inequality for the Kummer function of the phase shift of two parameters, Journal of Mathematical Sciences 178 (2) (2011) 178–186.
[15] D. Karp, S.M. Sitnik, Log-convexity and log-concavity of hypergeometric-like functions, Journal of Mathematical Analysis and Applications 364 (2010) 384–394.
[16] K.V. Mardia, P. Jupp, Directional Statistics, second ed., John Wiley & Sons, 2000.
[17] P.T. Spellman, G. Sherlock, M. Zhang, V.R. Iyer, K. Anders, M. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle regulated gene of the yeast Saccharomyces Cerevisia by microarray hybridization, Molecular Biology of the Cell 9 (1998) 3273–3297.
[18] S. Sra, Matrix nearness problems in data mining, Ph.D. Thesis, University of Texas at Austin, 2007.
[19] S. Sra, D. Karp, The multivariate Watson distribution: Maximum-likelihood estimation and other aspects, 2011. arXiv:stat.CO-1104.4422.
[20] A. Tanabe, K. Fukumizu, S. Oba, T. Takenouchi, S. Ishii, Parameter estimation for von Mises–Fisher distributions, Computational Statistics 22 (1) (2007) 145–157.
[21] G.S. Watson, Equatorial distributions on a sphere, Biometrika 52 (1-2) (1965) 193–201.