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)]

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

and $a^{\bar{0}} = 1$, $a^{\bar{j}} = a(a+1)\cdots(a+j-1)$, $j \ge 1$, 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 $(\boldsymbol{Q}\boldsymbol{\mu})^\top \boldsymbol{Q}\boldsymbol{x} = \boldsymbol{\mu}^\top \boldsymbol{x}$ for any orthogonal matrix $\boldsymbol{Q}$. In particular for $\boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{\mu}$, $\boldsymbol{\mu}^\top(\boldsymbol{Q}\boldsymbol{x}) = \boldsymbol{\mu}^\top \boldsymbol{x}$; 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

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

where $\boldsymbol{S} = n^{-1} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top$ 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

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