Generalized Watson Distribution on the Hypersphere with Applications to Clustering

Stephen J. Maybank* Liu Liu† Dacheng Tao‡

1 June 2022

Abstract. A family of probability density functions (pdfs) is defined on the unit hypersphere S^n. The parameter space for the pdfs is $G(d, n+1) \times \mathbb{R}_{\geq 0}$, for $1 \leq d \leq n$, where $G(d, n+1)$ is the Grassmannian of d dimensional linear subspaces in $\mathbb{R}^{n+1}$ and $\mathbb{R}_{\geq 0}$ is the range of values for a concentration parameter. This family of pdfs generalizes the Watson distribution on the sphere S^2. It is shown that the pdfs are tractable, in that *i)* a given pdf can be sampled efficiently, *ii)* the parameters of a pdf can be estimated using maximum likelihood, and *iii)* the Kullback-Leibler divergence and the Fisher-Rao metric on $G(d, n+1) \times \mathbb{R}_{\geq 0}$ have simple forms. A wide range of shapes of the pdfs can be obtained by varying d and the concentration parameter.

The pdfs are used to model clusters of feature vectors on the hypersphere. The clusters are compared using the Kullback-Leibler divergences of the associated pdfs. Experiments with the *mnist*, *Human Activity Recognition* and *Gas Sensor Array Drift* datasets show that good results can be obtained from clustering algorithms based on the Kullback-Leibler divergence, even if the dimension n of the hypersphere is high.

Keywords: classification; Fisher-Rao metric; generalised Watson distribution; Grassmannian; hypergeometric function; hypersphere; Kullback-Leibler divergence.

1 Introduction

The modelling of clusters of vectors is a fundamental task in data analysis. In many cases the vectors in a cluster have a common source which can form a basis for classification. In the probabilistic approach to clustering the models are probability density functions (pdfs) chosen from a family of pdfs which is parameterised by the points in a manifold. Each point θ in the manifold defines a unique pdf in the family of pdfs. Each cluster of vectors has an associated pdf θ that summarises the essential properties of the cluster. The pdfs are used to compare the different clusters. For example, if two clusters are modeled by similar pdfs θ_1 and θ_2, then this suggests that the clusters should be merged.

*Corresponding author, Department of Computer Science and Information Systems, Birkbeck College, Malet Street, London, WC1E 7HX, UK, steve.maybank@bbk.ac.uk

†School of Computer Science, J12 - School of Computer Science Building, The University of Sydney, NSW 2006, Australia, liu.liu1@sydney.edu.au

‡School of Computer Science, J12 - School of Computer Science Building, The University of Sydney, NSW 2006, Australia, dacheng.tao@gmail.com

If a new data vector is obtained, then it can be assigned to the cluster with a pdf that has the highest value of the likelihood, given the vector.

In applications, it is essential that the time complexity for estimating the parameter for a cluster and the time complexity for the calculation of the similarity between two parameter values should both be low. In addition, the measure of the similarity between two parameter values should be meaningful. In particular, the measure should be independent of the choice of parameterisation of the manifold. If ψ is an alternative parameterisation, then the similarity calculated using the parameters θ_1 and θ_2 should be the same as the similarity calculated using the parameters $\psi(\theta_1)$ and $\psi(\theta_2)$.

The natural candidate for measuring the similarity between θ_1 and θ_2 is the Kullback-Leibler divergence $D(\theta_1\|\theta_2)$ (Amari 1985; Cover & Thomas 2006). The abbreviation KL divergence is used consistently from this point onwards. It is straightforward to show that $D(\theta_1\|\theta_2) = D(\psi(\theta_1)\|\psi(\theta_2))$. However, there are relatively few parameterised families of pdfs for which the Kullback-Leibler divergence can be calculated with a low time complexity.

Many datasets consist of vectors in a high dimensional Euclidean space such that the directions of the vectors are more important than their lengths. Examples of such data include text (Dhillon & Modha 2001; Banerjee et al. 2005a; Hamsici & Martinez 2007), gene expression (Banerjee et al. 2005a; Hamsici & Martinez 2007) and face verification (Wang et al. 2017).The non-zero vectors in the Euclidean space $\mathbb{R}^{n+1}$ can be scaled to obtain vectors in the unit hypersphere S^n centred at the origin of $\mathbb{R}^{n+1}$. Pewsey & Garcia-Portugués (2021) provide a detailed review of the different types of data on the unit hypersphere, with a wide range of applications, including preshapes and gait analysis.

A new family of pdfs for modelling clusters of vectors on the unit hypersphere is defined in Section 3.2 below. The pdfs are generalisations of the Watson distribution (GWD) on S^2. In applications to classification the vectors in a cluster are assumed to be in or near to the intersection of S^n with a linear subspace L in $\mathbb{R}^{n+1}$. The components of the vectors normal to L have a Gaussian penalty. The GWD has a low time complexity, suitable for practical applications. The parameters defining L are estimated by using the singular value decomposition to find the maximum likelihood. The GWD has an additional scalar parameter, the concentration, which is estimated using a single implicit equation. The similarity of two given pdfs is measured using the KL divergence, which has a simple form depending on n, the dimension d of the two subspaces L_1 and L_2, the two concentration parameters and a single scalar parameter, $\mathrm{tr}(A)$ in (20), which is calculated from L_1 and L_2.

The dimension d of the linear subspace L is chosen to fit the application. If d is too small, then the concentration is likely to be low. Conversely, if d is too large, then the concentration is likely to be large.

1.1 Overview

Related work on the clustering of vectors in high dimensional hyperspheres is described in Section 2. The pdfs for the GWD are defined in Section 3 and an algorithm for parameter estimation based on maximum likelihood is defined. An efficient method for sampling from a pdf for the GWD is described. The KL divergence is defined in Section 4. The expression for the KL divergence is simplified, a symmetrised version of the Kullback-

Leibler divergence is defined and an expression for the Fisher-Rao metric is obtained.

Seven clustering algorithms are defined in Section 5. The algorithms include Kmeans (MacKay 2005) and Spkmeans (Dhillon & Modha 2001). Three of the remaining five algorithms use a symmetrised version of the KL divergence to measure the similarity of the model pdfs for any two given clusters. If the symmetrized KL divergence is small then the relevant clusters may be merged. The sixth algorithm uses maximum likelihood to fit pdfs to clusters. The final algorithm uses a mixture of von Mises-Fisher distributions to cluster data.

In Section 6 the clustering algorithms are tested on three databases, namely *mnist*[1] (LeCun et al. 1998), *Human Activity Recognition*[2] (Anguita et al. 2013) and *Gas Sensor Array Drift*[3] (Vergara et al. 2012). The accuracy of the results is discussed using the *normalised mutual information* (NMI). The best results are obtained from an algorithm (LSC-KL III) that uses the symmetrised KL divergence together with multiple initialisations to prevent the vectors forming unchangeable but erroneous clusters.

1.2 Appendices

There are four appendices. Appendix A contains the calculations necessary for the simplification of the KL divergence in Section 4. Appendix B contains additional calculations that are required to simplify the KL divergence still further and to obtain expressions for relevant quantities such as the scale factor for the GWD pdfs. Appendix C contains estimates of certain terms. The estimates are accurate if the concentration parameter is large. Appendix D reports experiments that use the Accuracy Rate or the Rand Index for classification in place of the NMI.

1.3 Supplement

The database mnist contains images of hand drawn digits in the range 0 to 9. A pdf is estimated for each digit i. The estimated pdf is then sampled. The supplement contains 30 sampled images, with three images for each value of i. It can be seen that the ith sampled image contains some of the features of the ith digit.

2 Related Work

Detailed surveys of probabilistic models for data on the hypersphere can be found in Chikuse (2003), Ley & Verdebout (2017), Ley & Verdebout (2018), Mardia & Jupp (1999), Mardia (1975) and Pewsey & Garcia-Portugués (2021). The last named survey is up to date and thorough. It includes a survey of the relevant publicly available software.

Section 2.1 describes related work on pdfs defined on the hypersphere. Section 2.2 describes related work on clustering.

[1] yann.lecun.com/exdb/mnist

[2] http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

[3] http://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset

2.1 Pdfs on the hypersphere

This subsection briefly reviews related work on families of pdfs defined on the hypersphere S^n for arbitary values of $n \geq 2$. In the definitions of the pdfs it is assumed that the measure for integration, $d\omega_n$, is induced on S^n by the Lebesgue measure on the Euclidean space $\mathbb{R}^{n+1}$. The terms c_{vmf}, c_{rs}, c_{tmf}, etc. in the expressions for the pdfs are scale factors chosen such that the integral of the pdf in question over S^n is unity.

2.1.1 Von Mises-Fisher distribution.

Let μ be a vector in S^n and let κ be a non-negative real number. The von Mises-Fisher distribution is defined by

$$p(x|\mu,\kappa)\, d\omega_n = c_{vmf} \exp\left(\kappa \mu^\top x\right)\, d\omega_n, \quad x \in S^n,$$

in which c_{vmf} is the scale factor and κ is a concentration parameter. Banerjee et al. (2005a) define an expectation maximisation (EM) algorithm for clustering vectors on S^n using a mixture of von Mises-Fisher distributions. Experiments on computer generated data, text data and gene expressions are reported.

2.1.2 Rotationally symmetric distributions

Let μ be a vector in S^n and let g be a function from $[-1,1]$ to $[0,\infty)$. The pdf

$$p(x|\mu,g)\, d\omega_n = c_{rs} g(\mu^\top x)\, d\omega_n, \quad x \in S^n$$

is said to be rotationally symmetric (Garcia-Portugués et al. 2020). For example, the von Mises-Fisher distribution is rotationally symmetric, with $g(t) = \exp(\kappa t)$, $\kappa \geq 0$. Garcia-Portugués et al. (2020) define two tests for rotational symmetry, given the value of μ.

2.1.3 Tangent distributions

Garcia-Portugués et al. (2020) define two extensions of the rotationally symmetric distributions, namely the tangent von Mises-Fisher distribution and the tangent elliptical distribution. Let μ be a vector in S^n and let Γ_μ be an $(n+1) \times n$ matrix such that the columns are an orthonormal basis of the orthogonal complement of μ. Let $u_\mu(x)$ be defined by

$$u_\mu(x) = \|\Gamma_\mu^\top x\|^{-1} \Gamma_\mu^\top x, \quad x \in S^n,$$

where $\|.\|$ is the Euclidean norm. Let κ be a non-negative real number and let ν be a vector in S^{n-1}. The tangent von Mises-Fisher distribution has the pdf obtained by Garcia-Portugués et al. (2020) in their Theorem 2,

$$p(x|\mu,g)\, d\omega_n = c_{tmf} g(\mu^\top x) \exp\left(\kappa \nu^\top u_\mu(x)\right)\, d\omega_n, \quad x \in S^n.$$

Let Λ be an $n \times n$ symmetric positive definite matrix such that the trace $\mathrm{tr}(\Lambda)$ of Λ equals n. The tangent elliptical distribution has the pdf obtained by Garcia-Portugués et al. (2020) in their Theorem 1,

$$p(\mu,g,\Lambda)\, d\omega_n = c_{te} g(\mu^\top x) \left(u_\mu^\top(x) \Lambda^{-1} u_\mu(x)\right)^{-n/2}\, d\omega_n, \quad x \in S^n.$$

These two tangent distributions are alternatives in tests for rotational symmetry, given that μ is unknown (Garcia-Portugués et al. 2020).

2.1.4 Spherical homoscedastic distributions

Two distributions on S^n are said to be spherically homoscedastic if their Bayes decision boundary is given by one or more hyperplanes. Hansici & Martinez (2007) show that the decision boundary for two spherically homoscedastic distributions coincides with the decision boundary for the Gaussian approximations to the two distributions.

Two von Mises-Fisher distributions with parameters μ_1, κ_1 and μ_2, κ_2 are spherically homoscedastic if $\kappa_1 = \kappa_2$ and $\mu_2 = R^\top \mu_1$, where R is in $SO(n+1)$.

Let A be a symmetric $(n+1) \times (n+1)$ matrix. The Bingham distribution on S^n (Bingham 1974) is defined by

$$p(x|A)\, d\omega_n = c_b \exp(x^\top A x)\, d\omega_n, \quad x \in S^n. \tag{1}$$

Two Bingham distributions with parameters A_1 and A_2 are spherically homoscedastic if $A_2 = R^\top A_1 R$, where R is the rotation of a plane spanned by any two of the eigenvectors of A_1. Full details are given by Hamsici & Martinez (2007).

2.1.5 Scaled von Mises-Fisher distribution

Let $a \in \mathbb{R}^{n+1}$ be a vector with components a_i such that $a_i > 0$, $1 \leq i \leq n+1$ and such that the product of the a_i is unity. Let $h(x, a)$ be the function defined by

$$h(x, a) = \sum_{j=1}^{n+1} (x_j / a_j)^2, \quad x \in S^n,$$

and let $T_a : S^n \to S^n$ be the transformation defined by

$$T_a(x) = h(x, a)^{-1/2} (a_1^{-1} x_1, \ldots, a_{n+1}^{-1} x_{n+1})^\top, \quad x \in S^n.$$

The transformation T_a is continuous and invertible. Let $T_a x$ have the von Mises-Fisher distribution with concentration parameter κ. Then the pdf for x is

$$p(x|T_a, \kappa)\, d\omega_n = c_s h(x, a)^{-n/2} \exp\left(h(a, x)^{-1/2} \kappa x_1 / a_1\right)\, d\omega_n, \quad x \in S^n.$$

The pdf $p(x|T_a, \kappa)$ is discussed in detail by Scealy & Wood (2019).

2.1.6 Elliptically symmetric angular Gaussian distribution

Let $\Phi(x|\mu, C)$ be the probability density function for the Gaussian distribution in $\mathbb{R}^{n+1}$ with expected value μ and covariance C, such that $C\mu = \mu$ and $\det(C) = 1$. The elliptically symmetric angular Gaussian distribution (ESAG) is defined by

$$p(x|\mu, C)\, d\omega_n = \left(\int_0^\infty \Phi(rx, \mu, C) r^n dr \right) d\omega_n, \quad x \in S^n.$$

Paine et al. (2017) show that the ESAG is tractable, in that the distributions can be simulated and the likelihood function can be computed, in both cases with a low time complexity.

2.2 Clustering on the hypersphere

Banerjee et al. (2005b) note that the cosine similarity measure of pairs of vectors yields good results for the classification of text documents. The cosine similarity depends on the directions of the two vectors but is independent of their lengths. In effect, the vectors are scaled to unit length. Banerjee et al. (2005a) describe a method for clustering vectors on the unit hypersphere based on von Mises-Fisher distributions. An expectation maximisation algorithm is used to find the clusters.

Dhillon & Modha (2001) describe the spherical Kmeans algorithm (Spkmeans) for clustering vectors on a hypersphere and apply it to the clustering of text documents. The dimension of the feature vectors may exceed 1000. Each cluster has a central concept vector. The linear subspace spanned by the concept vectors is found and the feature vectors are projected into it.

Hamsici & Martinez (2007) show that in some special cases the Bayes decision boundary for two clusters on a hypersphere S^n is also the Bayes decision boundary when the same clusters are modelled by Gaussian distributions defined on $\mathbb{R}^{n+1}$, as noted in Section 2.1.4. The hypersphere model for the data and the Gaussian model in $\mathbb{R}^{n+1}$ for the same data yield the same results for the classification of test vectors.

Zhao & Song (2018) use a heat kernel to measure the similarity between pairs of vectors on a hypersphere. The vectors are clustered using the heat kernel and a support vector machine. Yang et al. (2019) use the KL divergence in a hierarchical clustering method applied to geological data.

3 Generalised Watson Distribution

The parameter space for the pdfs in the generalised Watson distribution (GWD) is $G(d, n+1) \times \mathbb{R}_{\geq 0}$, where $G(d, n+1)$ is the Grassmann manifold for d-dimensional subspaces in $\mathbb{R}^{n+1}$ and $\mathbb{R}_{\geq 0}$ is the set of non-negative real numbers. The relevant properties of the Grassmann manifold are briefly reviewed in Section 3.1. The family of pdfs of the GWD is defined in Section 3.2. The action of the group of orthogonal matrices on the family of pdfs is described in Section 3.3. A maximum likelihood method for parameter estimation is described in Section 3.4. An efficient algorithm for sampling from the pdfs is summarised in Section 3.5.

3.1 Grassmann manifold

Information about the Grassmann manifold (Grassmannian) can be found in Chikuse (2003) and in Zhang et al. (2018). The Grassmann manifold $G(d, n+1)$ of d dimensional linear spaces in $\mathbb{R}^{n+1}$ is smooth and has dimension $d(n-d+1)$. Each d-dimensional subspace L of $\mathbb{R}^{n+1}$ is uniquely determined by the orthogonal projection matrix P with range L. Conversely, L uniquely determines P. The elements of $G(d, n+1)$ will be referred to as projection matrices or as d-dimensional subspaces, depending on the context.

An $(n+1) \times (n+1)$ matrix P is an orthogonal projection of $\mathbb{R}^{n+1}$ onto a subspace of dimension d if and only if P is symmetric, $P^2 = P$ and P has rank d. Let U be an $(n+1) \times (n+1)$ orthogonal matrix. The function $P \mapsto U^\top P U$ is a diffeomorphism of $G(d, n+1)$.

3.2 Definition of the GWD

The unit hypersphere S^n is given the measure $d\omega_n$ induced on it by the Lebesgue measure in $\mathbb{R}^{n+1}$. Let I be the $(n+1) \times (n+1)$ identity matrix and let P be a projection matrix in $G(d, n+1)$. The pdf $p(x|P, \kappa)$ for the GWD, conditional on P and the concentration parameter κ, is defined for x in S^n by

$$p(x|P, \kappa)\, d\omega_n = C_n(P, \kappa) \exp(-(\kappa/2)\|(I - P)x\|^2)\, d\omega_n, \quad x \in S^n, \kappa \in \mathbb{R}_{\geq 0}, \tag{2}$$

where $\|.\|$ is the Euclidean norm. The term $C_n(P, \kappa)$ is a normalising factor to ensure that the integral of $p(x|P, \kappa)$ over S^n is unity. The parameter θ discussed in Section 1 is in this case the pair (P, κ). Let L be the d dimensional subspace corresponding to P. The pdf (2) is constant for x in $L \cap S^n$ and this constant value is the maximum value of the pdf on S^n. If the concentration parameter κ is large, then $p(x|P, \kappa)$ is concentrated near to $L \cap S^n$. If κ is small, then $p(x|P, \kappa)$ approaches a uniform density on S^n.

The pdfs of the GWD are contained in the family of Bingham pdfs (1). If $\kappa \neq 0$ and if $A = -(\kappa/2)(I - P) + \lambda I$ in (1) for any given value of λ, then the pdf (2) is obtained. If P has rank 1 or rank n, then (2) reduces to a Watson distribution (Sra 2018; Sra & Karp 2013). If $n = 2$ and the projection P has rank 2 then the pdf (2) is similar to the girdle distribution defined by Selby (1964) in that the pdf is constant and a maximum on the great circle formed by the intersection of S^2 with the plane formed by the image of P.

The advantages of the generalised Watson distribution are *i)* the pdfs are much more computationally tractable than the pdfs for a general Bingham distribution; *ii)* a wide range of different forms of the pdfs are available as d, n and κ vary, *iii)* the KL divergence can be calculated in closed form, *iv)* if κ is large then the maximum likelihood values of the parameters P and κ can be estimated accurately using the well known singular value decomposition and some elementary arithmetic.

3.3 Action of the orthogonal group

Let U be an $(n+1) \times (n+1)$ orthogonal matrix. It follows from (2) that

$$p(x|U^\top PU, \kappa) = C_n(U^\top PU, \kappa) C_n(P, \kappa)^{-1} p(Ux|P, \kappa). \tag{3}$$

The integral of $p(Ux|P, \kappa)$ over S^n is equal to 1 because the measure $d\omega_n$ is invariant under the action of the orthogonal matrix U. On integrating (3) over S^n using the measure $d\omega_n$, it follows that

$$\begin{aligned} 1 &= \int_{S^n} p(x|U^\top PU, \kappa)\, d\omega_n, \\ &= C_n(U^\top PU, \kappa) C_n(P, \kappa)^{-1} \int_{S^n} p(Ux|P, \kappa)\, d\omega_n, \\ &= C_n(U^\top PU, \kappa) C_n(P, \kappa)^{-1}. \end{aligned} \tag{4}$$

If P is fixed, then any projection matrix in $G(d, n+1)$ can be expressed in the form $U^\top PU$ for an appropriate choice of U, thus it follows from (4) that $C_n(P, \kappa)$ is independent of P. Equation (3) then yields

$$p(x|U^\top PU, \kappa) = p(Ux|P, \kappa), \quad x \in S^n. \tag{5}$$

The notation $C_n(P,\kappa)$ is replaced by $C_n(\kappa)$.

If $\kappa_1 < \kappa_2$, then

$$\exp(-(\kappa_1/2)\|(I-P)x\|^2) \geq \exp(-(\kappa_2/2)\|(I-P)x\|^2).$$

It follows that $C_n(\kappa_1) \leq C_n(\kappa_2)$, thus

$$\frac{\partial C_n}{\partial \kappa} \geq 0, \quad \kappa > 0. \tag{6}$$

3.4 Parameter estimation

Let $x(i)$ for $1 \leq i \leq N$, be a set of N samples on S^n. A maximum likelihood algorithm for estimating the parameters of the GWD pdf $p(x|P,\kappa)$ is described. Similar calculations for the von Mises-Fisher distribution are described by Banerjee et al. (2005a), Hamsici & Martinez (2007) and Sra (2018).

Let X be an $(n+1)\times N$ data matrix with column vectors $x(i)$. It is assumed that the $x(i)$ are sampled independently from the pdf (2) with P and κ fixed. The log likelihood function for X is given by

$$L(X,P,\kappa) = N\ln(C_n(\kappa)) - \frac{1}{2}\kappa\sum_{i=1}^{N}\|(I-P)x(i)\|^2.$$

The log likelihood is maximised by first finding a projection matrix $\hat{P}$ that minimises

$$\sum_{i=1}^{N}\|(I-P)x(i)\|^2$$

and then finding the solution $\hat{\kappa}$ to the equation

$$N\frac{\partial}{\partial\kappa}\ln(C_n(\kappa)) = \frac{1}{2}\sum_{i=1}^{N}\|(I-\hat{P})x(i)\|^2, \quad \kappa \in \mathbb{R}_{\geq 0}. \tag{7}$$

The pair $\hat{P}$, $\hat{\kappa}$ are maximum likelihood estimates of the parameters P, κ in (2).

The projection matrix $\hat{P}$ can be obtained from the singular value decomposition of X. In detail, let the SVD of X be $X = U\Sigma V^\top$, where U is an $(n+1)\times(n+1)$ orthogonal matrix, Σ is a diagonal $(n+1)\times N$ matrix such that $\Sigma_{11} \geq \Sigma_{22} \geq \ldots$ and V is an $N\times N$ orthogonal matrix. It follows that

$$\begin{aligned}\sum_{i=1}^{N}\|(I-P)x(i)\|^2 &= \|(I-P)U\Sigma V^\top\|^2, \\ &= \|(I-U^\top PU)\Sigma\|^2,\end{aligned}$$

thus $I - U^\top\hat{P}U = D$, where D is the $(n+1)\times(n+1)$ projection matrix defined by

$$D_{ij} = \begin{cases} 1 & d+1 \leq i \leq n+1 \text{ and } j = i \\ 0 & \text{otherwise.}\end{cases}$$

It follows that $\hat{P} = U(I - D)U^\top$.

If the concentration parameter κ is not too large, then the implicit equation (7) can be solved for κ using standard numerical methods. If κ is large, let $R(\kappa)$ be the function

$$R(\kappa) = 2\frac{\partial}{\partial \kappa} \ln(C_n(\kappa)), \quad \kappa \geq 0. \tag{8}$$

It is shown in Appendix C that $R(\kappa)$ is approximated by

$$R(\kappa) = \kappa^{-1}(n - d + 1)(1 + O(\kappa^{-1})), \quad \kappa \gg 0. \tag{9}$$

It follows from (7), (8) and (9) that if the maximum likelihood value of κ is sufficiently large, then it is closely approximated by

$$\hat{\kappa} = (n - d + 1) \left(N^{-1} \sum_{i=1}^{N} \|(I - \hat{P})x(i)\|^2 \right)^{-1}.$$

The function $\kappa \mapsto R(\kappa)$ is investigated numerically in Fig. 1. The left-most column in Fig. 1 shows graphs of the function $\kappa \mapsto \kappa^{-1}(n - d + 1)$ together with sample points obtained from (8). The sample points are near to the relevant graph for $\kappa \geq 1000$. The middle column shows graphs of the function

$$\kappa \mapsto \ln(\text{abs}\left(R(\kappa) - \kappa^{-1}(n - d + 1)\right))$$

The graphs decrease rapidly for $\kappa \geq 500$.

Let $\tilde{R}(\kappa) = \kappa^{-1}(n - d + 1)$. Let $\Delta\kappa$ be a perturbation of κ such that $\tilde{R}(\kappa + \Delta\kappa)$ is equal to the true value $R(\kappa)$. A first order approximation to $\tilde{R}(\kappa)$ yields

$$\Delta\kappa = \text{abs}\left((R(\kappa) - \tilde{R}(\kappa))(d\tilde{R}(\kappa)/d\kappa)^{-1}\right). \tag{10}$$

The right-most column in Fig. 1 shows graphs of the function $\kappa \mapsto \Delta\kappa$, as defined by (10). The graphs tend to flatten for $\kappa \geq 2000$. This suggests that the fractional error, $\Delta\kappa/\kappa$ tends to zero as κ becomes large.

3.5 Sampling

Kent et al. (2018) describe an acceptance rejection method for sampling from any Bingham distribution, including the distributions defined by (2). However, a simpler method based on Saw (1978) is available for the generalised Watson distributions. It is shown in this subsection that a sample from (2) can be obtained by combining three samples from tractable distributions, namely two samples from hyperspheres with uniform densities and a sample from a distribution defined numerically on $[0, \pi/2]$.

In view of (5) it suffices to sample from the pdf $p(x|Q, \kappa)$ where the $(n + 1) \times (n + 1)$ projection matrix Q is defined by

$$Q_{ij} = \begin{cases} 1 & 1 \leq i \leq d \text{ and } j = i \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

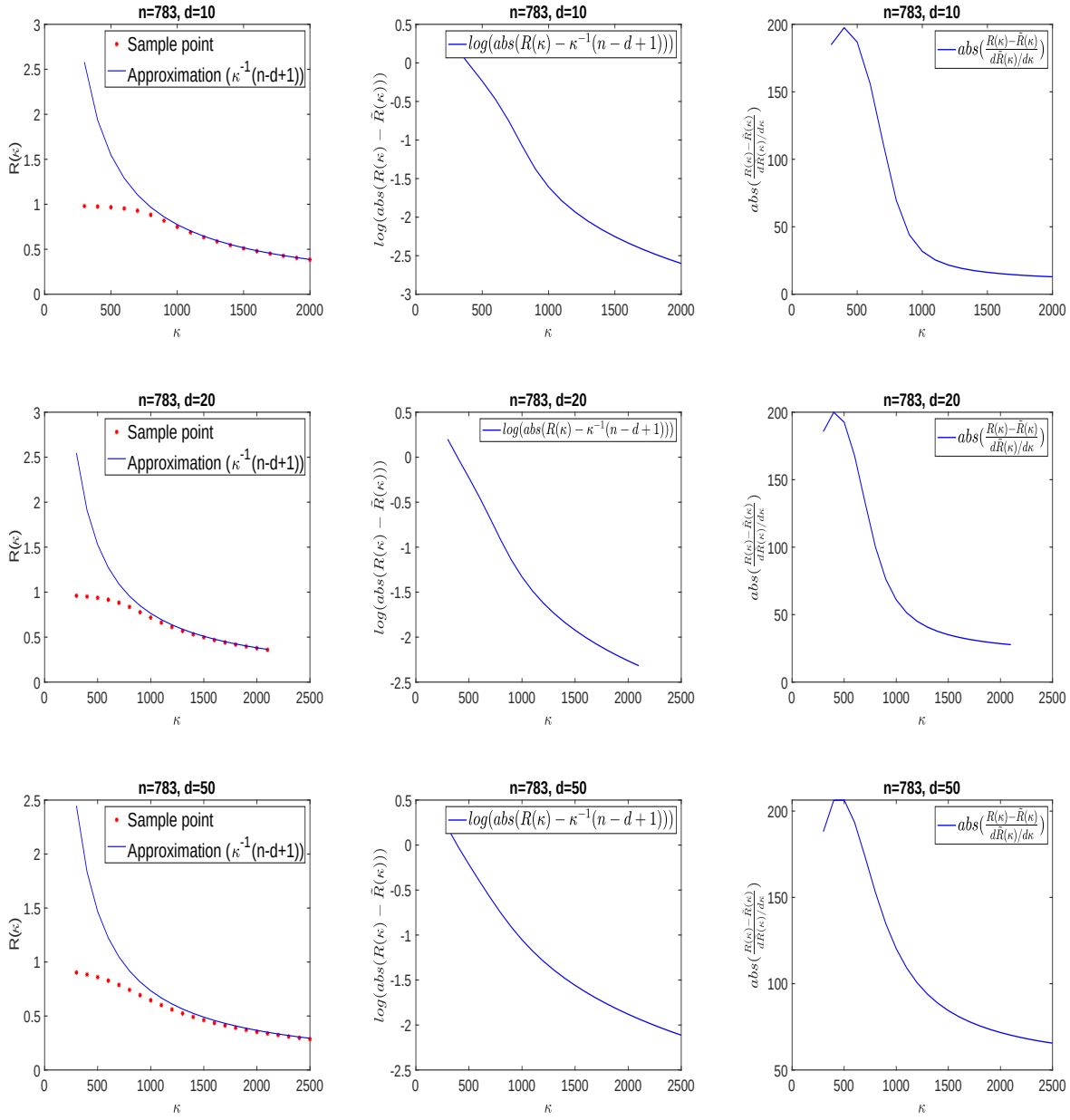


Figure 1: Numerical investigation of the function $\kappa \mapsto R(\kappa)$. The parameters are $n = 783$, and $d = 10, 20, 50$. First column: samples from (8) are compared with $\kappa \mapsto \kappa^{-1}(n-d+1)$. Second column: the log difference gap between $R(\kappa)$ and the function $\kappa \mapsto \kappa^{-1}(n-d+1)$. Third column: graphs of $\kappa \mapsto \text{abs}\left((R(\kappa) - \tilde{R}(\kappa))(d\tilde{R}(\kappa)/d\kappa)^{-1}\right)$.

Following Saw (1978), Section 1, let $q(\tilde{x}|Q,\kappa)$ be the pdf on $\mathbb{R}^{n+1}$ defined by

$$q(\tilde{x}|Q,\kappa)\,d\tilde{x} =$$

$$\left(2^{(n-1)/2}\Gamma((n+1)/2)\right)^{-1} p(\|\tilde{x}\|^{-1}\tilde{x}|Q,\kappa)\exp\left(-\|\tilde{x}\|^2/2\right)\,d\tilde{x},\quad \tilde{x}\in\mathbb{R}^{n+1}. \tag{12}$$

The notation $\tilde{x}$ is used in (12) to indicate a vector in $\mathbb{R}^{n+1}$. The notation x is reserved for vectors in S^n. On expressing $q(\tilde{x}|Q,\kappa)$ in polar coordinates, it is apparent that if s in $\mathbb{R}^{n+1}$ is a sample from $q(\tilde{x}|Q,\kappa)$, then $\|s\|$ and $\|s\|^{-1}s$ are independent and $\|s\|^{-1}s$ is a sample from $p(x|Q,\kappa)$.

It is shown that sampling from $q(\tilde{x}|Q,\kappa)$ is tractable. Let y in $\mathbb{R}^d$ and z in $\mathbb{R}^{n-d+1}$ be vectors such that $\tilde{x}=(y^\top,z^\top)^\top$. Let r_1, r_2 be defined by $r_1=\|y\|$, $r_2=\|z\|$ and let $d\omega_1$ and $d\omega_2$ be the Lebesgue measures in $\mathbb{R}^d$ and $\mathbb{R}^{n-d+1}$ respectively. It is noted that

$$d\tilde{x} = dydz = r_1^{d-1}r_2^{n-d}dr_1dr_2d\omega_1d\omega_2.$$

It follows that

$$q(\tilde{x}|Q,\kappa)d\tilde{x} = q(y,z,|Q,\kappa)r_1^{d-1}r_2^{n-d}dr_1dr_2d\omega_{d-1}d\omega_{n-d},\quad \tilde{x}\in\mathbb{R}^{n+1}.$$

In order to sample from $q(\tilde{x}|Q,\kappa)$ it is sufficient to sample points w_{d-1} from the hypersphere S^{d-1} and w_{n-d} from S^{n-d}, and to sample r_1, r_2 from the pdf

$$c\exp\left(-\kappa r_2^2(2(r_1^2+r_2^2))^{-1}-(r_1^2+r_2^2)/2\right)r_1^{d-1}r_2^{n-d}dr_1dr_2, \tag{13}$$

where c is a scale factor. The pdf (13) is simplified by setting $r_1=\sqrt{u}\cos(\phi)$, $r_2=\sqrt{u}\sin(\phi)$, to yield

$$2^{-1}c\exp\left(-\frac{1}{2}\kappa\sin^2(\phi)\right)\exp\left(-\frac{1}{2}u\right)\cos^{d-1}(\phi)\sin^{n-d}(\phi)u^{(n-1)/2}dud\phi. \tag{14}$$

It is apparent from (14) that u and ϕ are independent. The random variable u has a gamma distribution and ϕ has a distribution on $[0,\pi/2]$ which can be sampled numerically. In fact, the sample u is not required, because $p(x|P,\kappa)$ is defined only for unit vectors x.

4 Kullback-Leibler Divergence

The Kullback-Leibler divergence is a special case of the Bregman divergence (Banerjee et al. 2005b). It is the only divergence that has a minimisation property based on a mean value and that satisfies the data processing inequality (Painsky & Wornell 2019). The KL divergence for a given pair of pdfs is the expected value of the log likelihood ratio for the two pdfs (Amari 1985; Cover & Thomas 2006; MacKay 2005). It gives a way of comparing two pdfs that does not depend on the choice of parameterization. Kurz et al. (2016) establish a link between the KL divergence and moment matching for the von Mises and the Watson distributions.

An expression for the KL divergence from $p(x|P_2,\kappa_2)$ to $p(x|P_1,\kappa_1)$ is obtained in Section 4.1 and simplified in Section 4.2. Section 4.3 describes a symmetrical version of the KL divergence which is used in the clustering algorithms described in Sections 5. The Fisher-Rao metric associated with the KL divergence is obtained in Section 4.4. Background values for the KL divergence are discussed in Section 4.5.

4.1 Definition of the KL divergence

Let P be an orthogonal projection matrix in $G(d, n+1)$ and let E be an $(n+1) \times (n+1)$ symmetric matrix such that $P+E$ is in $G(d, n+1)$. The KL divergence $D(P, \kappa_1 \| P+E, \kappa_2)$ from $p(x|P+E, \kappa_2)$ to $p(x|P, \kappa_1)$ is defined by

$$D(P, \kappa_1 \| P+E, \kappa_2) = \int_{S^n} p(x|P, \kappa_1) \ln(p(x|P, \kappa_1)/p(x|P+E, \kappa_2))\, d\omega_n. \tag{15}$$

Let U be an $(n+1) \times (n+1)$ orthogonal matrix. It follows from (5) and (15) that the KL divergence is invariant under the action of U on $G(d, n+1)$, in that

$$D(U^\top P U, \kappa_1 \| U^\top (P+E) U, \kappa_2) =$$

$$\begin{aligned} &= \int_{S^n} p(x|U^\top P U, \kappa_1) \ln(p(x|U^\top P U, \kappa_1)/p(x|U^\top (P+E) U, \kappa_2))\, d\omega_n, \\ &= \int_{S^n} p(Ux|P, \kappa_1) \ln(p(Ux|P, \kappa_1)/p(Ux|P+E, \kappa_2))\, d\omega_n, \\ &= D(P, \kappa_1 \| P+E, \kappa_2). \end{aligned} \tag{16}$$

4.2 Expression for the KL divergence

An expression for the KL divergence $D(Q, \kappa_1 \| Q+E, \kappa_2)$ is obtained, where Q is the matrix defined by (11) and where E is an $(n+1) \times (n+1)$ matrix of the form

$$E = \begin{pmatrix} -A & C \\ C^\top & B \end{pmatrix} \tag{17}$$

such that $Q+E$ is in $G(d, n+1)$. The matrix A is of size $d \times d$, B is $(n-d+1) \times (n-d+1)$ and C is $d \times (n-d+1)$. The matrices E, A, B are symmetric. The KL divergence $D(P_1, \kappa_1 \| P_2, \kappa_2)$ for any two projection matrices P_1 and P_2 can be obtained from a KL divergence of the form $D(Q, \kappa_1 \| Q+E, \kappa_2)$ by choosing an orthogonal matrix U such that $U^\top P_1 U = Q$, noting that

$$D(P_1, \kappa_1 \| P_2, \kappa_2) = D(U^\top P_1 U, \kappa_1 \| U^\top P_2 U, \kappa_2) = D(Q, \kappa_1 \| Q + (U^\top P_2 U - Q), \kappa_2),$$

and setting $E = U^\top P_2 U - Q$.

Let x be a point on the hypersphere S^n and let y in $\mathbb{R}^d$ and z in $\mathbb{R}^{n-d+1}$ be points such that $x = (y^\top, z^\top)^\top$. Let $\langle . \rangle_1$ indicate integration over S^n with weight $p(x|Q, \kappa_1)$. It follows from (15) that

$$D(Q, \kappa_1 \| Q+E, \kappa_2) = \left\langle \ln \left(\frac{C_n(\kappa_1) \exp(-\kappa_1 \|z\|^2/2)}{C_n(\kappa_2) \exp(-\kappa_2 \|(I-Q-E)x\|^2/2)} \right) \right\rangle_1,$$

$$= \ln \left(\frac{C_n(\kappa_1)}{C_n(\kappa_2)} \right) - \frac{1}{2} \kappa_1 \langle \|z\|^2 \rangle_1 + \frac{1}{2} \kappa_2 (\langle \|z\|^2 \rangle_1 - 2\langle x^\top E(I-Q)x \rangle_1 + \langle \|Ex\|^2 \rangle_1). \tag{18}$$

It is convenient to define the function $g(d, n, \kappa_1)$ by

$$g(d, n, \kappa_1) = d^{-1} \langle \|y\|^2 \rangle_1 - (n-d+1)^{-1} \langle \|z\|^2 \rangle_1. \tag{19}$$

Expressions for $\langle x^\top E(I-Q)x\rangle_1$ and $\langle \|Ex\|^2\rangle_1$ are obtained in Appendix A. It follows from (33) and (35) in Appendix A, and from (18) and (19) that

$$D(Q,\kappa_1\|Q+E,\kappa_2) = \ln\left(C_n(\kappa_1)/C_n(\kappa_2)\right) + \frac{1}{2}(\kappa_2-\kappa_1)\langle\|z\|^2\rangle_1 + \frac{1}{2}\mathrm{tr}(A)\kappa_2 g(d,n,\kappa_1), \quad (20)$$

where $\mathrm{tr}(A)$ is the trace of the matrix A. It is apparent from (20) that the KL divergence depends on E only through the single scalar parameter $\mathrm{tr}(A)$. Expressions for $\langle\|z\|^2\rangle_1$ and $\langle\|y\|^2\rangle_1$ are given by the equations (48) and (49) respectively in Appendix B.

4.3 Symmetry

The KL divergence (15) is in general not symmetric, in that the value of the KL divergence may change if P, κ_1 and $P+E$, κ_2 in (15) are interchanged. However, the KL divergence is symmetric if the two pdfs have the same concentration parameter, $\kappa_1 = \kappa_2$. This symmetry is a consequence of the following result: let P_1 and P_2 be projections and let κ_1 and κ_2 be the corresponding concentrations. Then it follows that

$$D(P_1,\kappa_1\|P_2,\kappa_2) = D(P_2,\kappa_1||P_1,\kappa_2). \quad (21)$$

It suffices to consider the case $P_1 = Q$, $P_2 = Q+E$, where Q is defined by (11) and E is defined by (17). Let U be an orthogonal matrix such that $U^\top(Q+E)U = Q$. It follows from (16) that

$$D(Q+E,\kappa_1\|Q,\kappa_2) = D(Q,\kappa_1\|U^\top QU,\kappa_2) = D(Q,\kappa_1\|Q+(U^\top QU-Q),\kappa_2).$$

It follows from the definition of U that $E = UQU^\top - Q$. Let $\tilde{E}$ be the matrix defined by

$$\tilde{E} = U^\top QU - Q.$$

It suffices to prove that

$$D(Q,\kappa_1\|Q+\tilde{E},\kappa_2) = D(Q,\kappa_1\|Q+E,\kappa_2).$$

Let $\tilde{E}$ have the same block structure as E, with blocks $\tilde{A}$, $\tilde{B}$, $\tilde{C}$. It follows from (20) that it suffices to prove that $\mathrm{tr}(A) = \mathrm{tr}\left(\tilde{A}\right)$. Let the orthogonal matrix U have the block structure

$$U = \begin{pmatrix} U(1) & U(2) \\ U(3) & U(4) \end{pmatrix},$$

in which the $U(i)$ have the respective dimensions $d\times d$, $d\times(n-d+1)$, $(n-d+1)\times d$ and $(n-d+1)\times(n-d+1)$. It follows from the definitions of E and $\tilde{E}$ that

$$\begin{aligned} \mathrm{tr}(A) &= \mathrm{tr}(U(1)U(1)^\top) - d, \\ \mathrm{tr}\left(\tilde{A}\right) &= \mathrm{tr}(U(1)^\top U(1)) - d, \end{aligned}$$

thus $\mathrm{tr}(A) = \mathrm{tr}\left(\tilde{A}\right)$ as required.

It follows from (21) that the KL divergence is symmetric if $\kappa_1 = \kappa_2$.

4.4 Symmetrisation and the Fisher-Rao metric

In many applications it is convenient to use a symmetrised version of the KL divergence. The symmetrisation $D_S(Q, \kappa_1, Q+E, \kappa_2)$ is defined by

$$D_S(Q, \kappa_1, Q+E, \kappa_2) = \frac{1}{2}D(Q, \kappa_1 \| Q+E, \kappa_2) + \frac{1}{2}D(Q+E, \kappa_2 \| Q, \kappa_1).$$

Let $\langle . \rangle_i$ indicate integration over S^n with weight $p(x|Q, \kappa_i)$ for $i = 1, 2$. It follows from (20) and (21) that

$$D_S(Q, \kappa_1, Q+E, \kappa_2) =$$

$$\frac{1}{4}(\kappa_2 - \kappa_1)\left(\langle \|z\|^2 \rangle_1 - \langle \|z\|^2 \rangle_2\right) + \frac{1}{4}\mathrm{tr}(A)(\kappa_2 g(d, n, \kappa_1) + \kappa_1 g(d, n, \kappa_2)). \quad (22)$$

The first term on the right-hand side of (22) is approximated. Let $f(\kappa)$ be the function defined by

$$f(\kappa) \equiv -\frac{\partial^2}{\partial \kappa^2} \ln(C_n(\kappa)) = \left(\frac{C_n'(\kappa)}{C_n(\kappa)}\right)^2 - \frac{C_n''(\kappa)}{C_n(\kappa)}, \quad (23)$$

where $'$ indicates the partial derivative with respect to κ. A Taylor expansion of $\langle \|z\|^2 \rangle_2$ about κ_1 yields

$$\langle \|z\|^2 \rangle_2 = \langle \|z\|^2 \rangle_1 + (\kappa_2 - \kappa_1)\left(\frac{\partial}{\partial \kappa}\langle \|z\|^2 \rangle\right)_{\kappa=\kappa_1} + O((\kappa_2 - \kappa_1)^2). \quad (24)$$

It follows from (24) and (50) in Appendix B that

$$\begin{aligned} \langle \|z\|^2 \rangle_1 - \langle \|z\|^2 \rangle_2 &= -2(\kappa_2 - \kappa_1)\left(\frac{\partial^2}{\partial \kappa^2}\langle \|z\|^2 \rangle\right)_{\kappa=\kappa_1} + O((\kappa_2 - \kappa_1)^2) \\ &= 2(\kappa_2 - \kappa_1) f(\kappa_1) + O((\kappa_2 - \kappa_1)^2). \end{aligned} \quad (25)$$

It follows from (22) and (25) that

$$D_S(Q, \kappa_1, Q+E, \kappa_2) =$$

$$2^{-1}(\kappa_2 - \kappa_1)^2 f(\kappa_1) + 4^{-1}\mathrm{tr}(A)(\kappa_2 g(d, n, \kappa_1) + \kappa_1 g(d, n, \kappa_2)) + O((\kappa_2 - \kappa_1)^2). \quad (26)$$

Let θ be a parameter vector for $G(d, n+1)$ and let κ be a parameter for $\mathbb{R}_{\geq}$. The Fisher-Rao metric is given by a matrix $J(\theta, \kappa)$ such that

$$D(\theta, \kappa \| \theta + \Delta\theta, \kappa + \Delta\kappa) = 2^{-1}(\Delta\theta, \Delta\kappa)^\top J(\theta, \kappa)(\Delta\theta, \Delta\kappa) + O_3, \quad (27)$$

where O_3 consists of terms of third or higher order in $\Delta\theta$ and $\Delta\kappa$ (Amari 1985). Alternatively, the same Fisher-Rao metric can be obtained from

$$D(\theta + \Delta\theta, \kappa + \Delta\kappa \| \theta, \kappa) = 2^{-1}(\Delta\theta, \Delta\kappa)^\top J(\theta, \kappa)(\Delta\theta, \Delta\kappa) + O_3. \quad (28)$$

It follows from (27)) and (28) that the Fisher-Rao metric can be obtained from the symmetrised version of the KL divergence in (22),

$$D_S(\theta, \kappa, \theta + \Delta\theta, \kappa + \Delta\kappa) = 2^{-1}(\Delta\theta, \Delta\kappa)^\top J(\theta, \kappa)(\Delta\theta, \Delta\kappa) + O_3.$$

Let E be given by (17). The matrix $Q+E$ is a projection, thus $(Q+E)(Q+E) = (Q+E)$, from which it follows that

$$\begin{aligned} A &= CC^\top + O_4(E), \\ B &= C^\top C + O_4(E), \\ E &= \begin{pmatrix} -CC^\top & C \\ C^\top & C^\top C \end{pmatrix} + O_4(E), \end{aligned} \tag{29}$$

where the terms $O_4(E)$ are fourth order in the entries of E. See (34) in Appendix A.

Let $\theta(Q)$ be the value of θ at the point Q in $G(d, n+1)$. It follows from (26) and (29) that

$$D_S(\theta(Q), \kappa, \theta(Q)+\Delta\theta, \kappa+\Delta\kappa) = 2^{-1} f(\kappa)\Delta\kappa^2 + 2^{-1}\kappa g(d,n,\kappa)\|C\|^2 + O_3,$$

where $\|.\|$ is the Euclidean norm. The matrix C is flattened to give a $d(n-d+1)$ dimensional vector c. Set $\Delta\theta_i = c_i$, $1 \le i \le d(n-d+1)$. The vector $\theta(Q)+\Delta\theta$ is an approximation to $\theta(Q+E)$ with an error of order four in the entries of E. The matrix $J(\theta(Q), \kappa)$ is given by

$$\begin{aligned} J_{ii}(\theta(Q),\kappa) &= \kappa g(d,n,\kappa), \quad 1 \le i \le d(n-d+1), \\ J_{ii}(\theta(Q),\kappa) &= f(\kappa), \qquad\qquad i = d(n-d+1)+1, \\ J_{ij}(\theta(Q),\kappa) &= 0, \qquad\qquad 1 \le i,j \le d(n-d+1)+1, i \ne j. \end{aligned}$$

The Fisher-Rao metric defines a distribution on $G(d, n+1) \times \mathbb{R}_{\ge}$ suitable for Bayesian parameter estimation.

4.5 Background KL divergence

A background value of the KL divergence is obtained by sampling $Q+E$ in (20) from the uniform density on $G(d, n+1)$. If an experimental value, $D(Q, \kappa_1 \| Q+E, \kappa_2)$, of the KL divergence is less than the background value, then this indicates that the two clusters modelled by (Q, κ_1) and $(Q+E, \kappa_2)$ are connected in some way and could be candidates for merging. The uniform distribution on $G(d, n+1)$ is obtained by scaling the Fisher-Rao measure on $G(d, n+1)$ such that the resulting volume of $G(d, n+1)$ is unity (James 1954).

If d and n are sufficiently large, then the background value of $\mathrm{tr}(A)$ in (20) can be approximated by a function of d and n with only a small error. In detail, let Z be a random $(n+1) \times d$ matrix such that the entries of Z are sampled independently from a Gaussian distribution with expected value 0 and variance 1. Let M be the $(n+1) \times d$ matrix given by $M = Z(Z^\top Z)^{-1/2}$. It is shown in Chikuse (2003) that $P \equiv MM^\top$ is a sample from the uniform distribution on $G(d, n+1)$. The KL divergence, $D(Q, \kappa_1 \| Q+E, \kappa_2)$, depends on the projection P only through the term $\mathrm{tr}(A)$ in (20), where A is the $d \times d$ matrix defined in (17).

It follows from (17) and the condition $P = Q+E$ that $A_{ij} = Q_{ij} - P_{ij}$, $1 \le i,j \le d$, thus

$$\mathrm{tr}(A) = d - \sum_{i,j=1}^{d} M_{ij}^2.$$

If n is large, then $(Z^\top Z)^{-1/2}$ is approximated by a $d \times d$ diagonal matrix with entries $(n+1)^{-1/2}$ on the diagonal. It follows that M is approximated by $(n+1)^{-1/2}Z$. If in addition, d is large, then

$$\mathrm{tr}(A) \approx d - \sum_{i,j=1}^{d} (n+1)^{-1} = d - d^2/(n+1). \tag{30}$$

If $\kappa_1 = \kappa_2 = \kappa$, then it follows from (20) and (30) that

$$D(Q, \kappa \| Q + E, \kappa) \approx 2^{-1}(d - d^2(n+1)^{-1})\kappa g(d, n, \kappa).$$

Any two clusters with a KL divergence significantly less that the background KL divergence are candidates for merging.

Experiments with computer generated data confirm that the approximation (30) to $\mathrm{tr}(A)$ is accurate if d and n are large. For example, if $d = 10$ and $n = 100$, then $d - d^2(n+1)^{-1} = 9.010$. A set of ten random samples of P yields sample values for $\mathrm{tr}(A)$ with an expected value of 9.000 and a standard deviation of 0.127. Experiments also show that values of the KL divergence are similar for different samples of $P = Q + E$. For example, if $n = 100$, $d = 10$ and $\kappa = 50$, then the values of $D(Q, \kappa \| Q + E, \kappa)$ for 100 random samples of P have a mean of 1.664 and a standard deviation of 0.023.

5 Clustering

In this section and the next the effectiveness of the generalised Watson distribution for clustering vectors on a hypersphere is assessed experimentally. Four of the clustering algorithms are based on linear subspaces in $\mathbb{R}^{n+1}$. The first algorithm is probability-based (Algorithm 1). The remaining three algorithms are parameter-based (Algorithms 2, 3 and 4). In the description of the algorithms given below the feature vectors $x(i)$, $1 \leq i \leq N$, form the columns of an $(n+1) \times N$ matrix X. The vectors in a given cluster i form the columns of a matrix X_i. The required number of clusters is c and the subspace associated with each cluster has dimension d.

The four linear subspace based algorithms are described in Sections 5.1 to 5.4 respectively. The experimental results obtained for these four algorithms are described in Section 6.

5.1 Linear subspace clustering based on probability

Algorithm 1 (LSC-HS(d,K)) finds an initial estimate of the clusters using Kmeans (MacKay 2005) or Spkmeans (Dhillon & Modha 2001). This initialisation is followed by alternating the steps maximization and expectation. Let c be a specified number of clusters. The maximization step yields the estimates of the parameters $\hat{\theta}_h$,

$$\hat{\theta}_h = \left\{ \hat{P}_h, \hat{\kappa}_h \right\}, \quad 1 \leq h \leq c.$$

The estimate $\hat{P}_h$ of the projection matrix is obtained using the singular value decomposition, as described in Section 3.4. The estimate $\hat{\kappa}_h$ of the concentration is obtained by

Algorithm 1 LSC-HS (d, c)

Require: the matrix X with columns $x(i)$ for $1 \leq i \leq N$, the number c of clusters and the dimension d of the subspaces. Initialise a c-partition $\{X_h\}_{h=1}^{c}$ using the Spkmeans or the Kmeans algorithms

for t=1 to T **do**
  Maximization Step
  for h=1 to c **do**
    Compute the SVD $X_h = U_h \Sigma_h V_h^\top$
    Compute the projection matrix $\hat{P}_h = U_h Q U_h^\top$
    Estimate the concentration parameter $\hat{\kappa}_h$.
  end for
  Expectation Step
  Set $X_h = \emptyset$ for $1 \leq h \leq c$.
  for i=1 to N **do**
    Normalize the sample to unit hypersphere, $x(i) = x(i)/\left\|x(i)\right\|$,
    Calculate the values of the probability density functions

$$p_h(x(i)|\hat{P}_h, \hat{\kappa}_h) = C_n\left(\hat{\kappa}_h\right) e^{-\hat{\kappa}_h \left\|\left(I-\hat{P}_h\right)x(i)\right\|^2/2}, 1 \leq h \leq c.$$

    Include $x(i)$ as a column of X_h, where $h = \text{argmax}_{h'}\ p\left(x(i)|\hat{P}_{h'}, \hat{\kappa}_{h'}\right)$.
  end for
  Delete any sets X_h that are empty, $1 \leq h \leq c$
  $c = |\{X_h, X_h \neq \emptyset, 1 \leq h \leq c\}|$
end for

solving (7) numerically, given $\hat{P}_h$. In the expectation step each sample vector is assigned to the cluster with the highest value of the pdf for the vector. The maximization step and the expectation step are computed iteratively until the estimates $\hat{P}_h$ and $\hat{\kappa}_h$ converge. The pseudo-code is given in the table for Algorithm 1.

5.2 Linear subspace clustering based on the KL divergence I

Algorithm 2 (LSC-KL I(d, c, c_0)) carries out linear subspace clustering based on the symmetrisation (22) of the KL divergence. The feature vectors $x(i)$ are first divided into c_0 disjoint clusters, where $c_0 > c$. Then, the symmetrized KL divergence is used to merge pairs of clusters until c clusters are obtained. The pseudo-code is given in the table for Algorithm 2.

5.3 Linear subspace clustering based on the KL divergence II

The vectors in cluster i form the columns of a matrix X_i as noted in the first paragraph of Section 5. The mean value of the vectors contributing to X_i is

$$\bar{x}(i) = \frac{1}{|X_i|} \sum\nolimits_{x(k) \in X_i} x(k),$$

Algorithm 2 LSC-KL I (d, c, c_0)

Require: the matrix X with columns $x(i)$ for $1 \leq i \leq N$, the number c of clusters, the dimension d of the subspaces and an integer $c_0 > c$

Initialize the set $\mathcal{A} = \{h, 1 \leq h \leq c_0\}$
Initialize the c_0-partition $\{X_h\}_{h=1}^{c_0}$ using LSC-HS(d, c_0) and estimate the parameters $\hat{\theta}_h = \left\{\hat{P}_h, \hat{\kappa}_h\right\}, \quad 1 \leq h \leq c_0$
repeat
    Find $(i', j') = \text{argmin}_{i,j \in \mathcal{A}, i<j} D_S(\hat{\theta}_i \| \hat{\theta}_j)$
    Merge $X_{i'}$, $X_{j'}$. Label the new cluster with i'. Delete j' from $\mathcal{A}$.
    Compute the parameters $\hat{\theta}_{i'}$ of the new cluster $X_{i'}$.
    $c_0 = c_0 - 1$
until $c_0 = c$

where $|X_i|$ is the number of columns in X_i. The distance $\text{dist}(i, j)$ is defined by

$$\text{dist}(i, j) = \|\bar{x}(i) - \bar{x}(j)\|,$$

where $\|.\|$ is the Euclidean norm. Algorithm 3 (LSC-KL II(d, c, c_0)) first divides the feature vectors $x(i)$ into c_0 distinct clusters, where $c_0 > c$. The algorithm then finds the set $\mathcal{B}$ of pairs (i, j) of clusters for which $i < j$ and $\text{dist}(i, j)$ is less than a given threshold δ. The symmetrised KL divergence is then minimized over $\mathcal{B}$. The pair of clusters (i', j') in $\mathcal{B}$ at which the symmetrised KL divergence has a minimum are merged. The merging of clusters is iterated until c clusters are obtained. The pseudo code is given in the table for Algorithm 3.

Algorithm 3 LSC-KL II (d, c, c_0)

Require: the matrix X with columns $x(i)$ for $1 \leq i \leq N$, the number c of clusters, the dimension d of the subspaces, an integer $c_0 > c$ and a threshold δ.

Initialize the set $\mathcal{A} = \{i, 1 \leq i \leq c_0\}$.
Initialize the c_0-partition $\{X_h\}_{h=1}^{c_0}$ using LSC-HS(d, c_0) and estimate the parameters $\hat{\theta}_h = \left\{\hat{P}_h, \hat{\kappa}_h\right\}, \quad 1 \leq h \leq c_0$
repeat
    $\delta_0 = \delta$
    repeat
        Set $\mathcal{B} = \{(i, j), 1 \leq i < j \leq c_0, \text{dist}(i, j) \leq \delta\}$
        $\delta = 1.1 * \delta$
    until $\mathcal{B}\,! = NULL$
    $\delta = \delta_0$
    $(i', j') = \text{argmin}_{(i,j) \in \mathcal{B}} D_S(\hat{\theta}_i \| \hat{\theta}_j)$
    Merge $X_{i'}$ and $X_{j'}$. Label the new cluster with i'. Delete j' from $\mathcal{A}$
    Compute the parameters $\hat{\theta}_{i'}$ of the new cluster $X_{i'}$
    $c_0 = c_0 - 1$
until $c_0 = c$

5.4 Linear subspace clustering based on the KL divergence III

In the algorithm LSC-HS the parameters are changed at each maximization step. However, if one cluster is well separated from the other clusters, then the parameters for that cluster do not change. This observation motivates LSC-KL III(d, c, c_0) in which the initialization is repeated for each iteration. The pseudo code is given in the table for Algorithm 4.

Algorithm 4 LSC-KL III (d, c, c_0)

Require: the matrix X with columns $x(i)$ for $1 \leq i \leq N$, the number c of clusters, the dimension d of the subspaces, an integer $c_0 > c$ and a threshold δ.

Initialize the set $\mathcal{A} = \{i, 1 \leq i \leq c_0\}$
$T = c_0 - c$
for $t = 1$ to T **do**
    Initialize the c_0-partition $\{X_h^{t-1}\}_{h=1}^{c_0}$ using LSC-HS(d, c_0)
    $\delta_0 = \delta$
    repeat
        Set $\mathcal{B} = \{(i, j), 1 \leq i < j \leq c_0, \text{dist}(i, j) \leq \delta\}$
        $\delta = 1.1 * \delta$
    until $\mathcal{B} != NULL$
    $\delta = \delta_0$
    $(i', j') = \text{argmin}_{(i,j)\in\mathcal{B}}\{D_S(\hat{\theta}_i \| \hat{\theta}_j)\}$
    Merge $X_{i'}^{t-1}$ and $X_{j'}^{t-1}$. Label the new cluster with i'. Delete j' from $\mathcal{A}$
    $c_0 = c_0 - 1$
end for

6 Experiments

Seven clustering algorithms are tested experimentally on the *mnist* (LeCun et al. 1998), *Human Activity Recognition* (HAR) (Anguita et al. 2013), and *Gas Sensor Array Drift* (GSAD) (Vergara et al. 2012) datasets. The seven algorithms comprise the four algorithms described in Section 5 and three algorithms taken from the literature. The algorithms are listed in Section 6.1. The data for the testing is described in Section 6.2 and the results are described in Section 6.3.

6.1 Algorithms

The following seven algorithms are compared to assess the quality of clustering: 1) Kmeans; 2) Spkmeans; 3) Mixture of von Mises-Fisher distributions (moVMF); 4) Algorithm 1: Linear subspace clustering using a hypersphere (LSC-HS); 5) Algorithm 2: Linear subspace clustering using the symmetrized KL divergence (LSC-KL I); 6) Algorithm 3: Linear subspace clustering using the symmetrized KL divergence and dist(i, j) (LSC-KL II); 7) Algorithm 4: Linear subspace clustering using the symmetrized KL divergence, dist(i, j) and LSC-HS for reinitialization at each stage (LSC-KL III).

The Kmeans algorithm is described by MacKay (2005) and the Spkmeans algorithm is described by Dhillon & Modha (2001). The mixture of von-Mises-Fisher distributions is described by Banerjee et al. (2005a), Banerjee et al. (2005b), and Sra (2016). The concentration parameters are estimated using the algorithm described in Section 3.4.

6.2 Datasets and methodology

In all cases, the feature vectors are normalized to obtain vectors on the unit hypersphere.

Mnist The mnist dataset consists of images of the digits 0 to 9. There are ten clusters in mnist, namely $\{\mathcal{C}_i\}_{i=0}^{9}$. The total number of images is 10,000. Each cluster $\mathcal{C}_i$ contains $N = 1,000$ images of the digit i for $0 \leq i \leq 9$. Four subsets of the data are chosen, namely, $\mathcal{D}_1 = \{\mathcal{C}_1, \mathcal{C}_3\}$, $\mathcal{D}_2 = \{\mathcal{C}_4, \mathcal{C}_5, \mathcal{C}_6\}$, $\mathcal{D}_3 = \{\mathcal{C}_0, \mathcal{C}_2, \mathcal{C}_4, \mathcal{C}_6, \mathcal{C}_8\}$, $\mathcal{D}_4 = \{\mathcal{C}_i\}_{i=0}^{9}$. The following parameter settings are employed.

- Dimension of the hypersphere: $n = 783$.
- d-dimensional subspaces: $d = 10$, 20, 50, 100.
- The initial number of clusters for Algorithms 2, 3, and 4: $c_0 = 2c$, where c is the number of clusters.
- Threshold values for Algorithms 3 and Algorithms 4: $\delta = 0.1, 0.3, 0.5$.

Human Activity Recognition The Human Activity Recognition (HAR) video dataset contains six types of human actions, namely WALKING, WALKING-UPSTAIRS, WALKING-DOWNSTAIRS, SITTING, STANDING and LAYING. Each person performed all six actions while wearing a Samsung Galaxy S II smartphone on the waist. There are 6 clusters, one for each activity. The total number of images is 7,415. There are approximately $N \approx 1,100$ samples in each class. The parameters' setting are the same as for mnist, except that the dimension of the hypersphere is $n = 560$ and the values of d are $d = 5, 10, 20, 40, 50$. For further information, see Anguita et al. (2013).

Gas Sensor Array Drift The Gas Sensor Array Drift (GSAD) dataset covers six distinct pure gaseous substances, namely Ammonia, Acetaldehyde, Acetone, Ethylene, Ethanol, and Toluene. The dataset was gathered during the period of January 2008 to February 2011 (36 months) in a gas delivery platform facility situated at the ChemoSignals Laboratory in the BioCircuits Institute (Vergara et al. 2012). There are 6 clusters, one for each gas. Each feature vector contains 8 components extracted from each of 16 sensors, resulting in a 128-dimensional feature vector (8 features $\times$ 16 sensors). The total number of samples is 6,000. There are approximately $N \approx 1,000$ images in each class. The parameters' setting is the same as the dataset mnist, except that the dimension of the hypersphere is $n = 127$ and the values of d are $d = 2, 4, 6, 8, 10, 15$. Each feature vector is centered to have 0 mean before scaling to obtain a vector on the hypersphere.

The performance of the algorithms is assessed using *normalized mutual information* (NMI) as a measure of the statistical similarity between a cluster and the ground truth (Strehl & Ghosh 2002; Vinh & Epps 2009; Vinh et al. 2010; Vinh et al. 2009). Let p_J be the joint probability for the ground truth cluster and the empirical cluster, let p_M and

Table 1: Comparison of NMI results for the datasets: $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$, d-dimension subspace: $d = 10, 20, 50, 100$, and the dimension of the hypersphere is $n = 783$.

Data	dim	kmeans	Spkmeans	moVMF	LSC-HS	LSC-KL I	LSC-KL II	LSC-KL III
$\mathcal{D}_1$	10	0.7807	0.7716	0.6902	0.9231	0.3736	0.8546	**0.9333**
	20				0.9248	0.3805	0.9237	**0.9402**
	50				**0.9245**	0.3302	0.7904	0.9086
	100				0.8168	0.2805	0.7169	**0.8439**
$\mathcal{D}_2$	10	0.4785	0.4793	0.4853	0.5975	0.6104	0.5973	**0.6197**
	20				0.6077	0.5226	**0.6174**	0.5868
	50				0.5405	0.4346	0.5512	**0.5608**
	100				0.4901	0.3394	0.4942	**0.5391**
$\mathcal{D}_3$	10	0.4744	0.4852	0.4988	0.5588	0.5503	0.4911	**0.6120**
	20				0.5236	0.4660	0.5225	**0.5925**
	50				**0.4923**	0.3888	0.4638	0.4598
	100				**0.4812**	0.3979	0.4349	0.4159
$\mathcal{D}_4$	10	0.5076	0.5109	0.1486	**0.5914**	0.4997	0.4999	0.5677
	20				0.5838	0.4731	0.4922	**0.6121**
	50				**0.5386**	0.4273	0.4725	0.5270
	100				**0.5146**	0.4300	0.4644	0.4731

Table 2: Comparison of NMI results for the datasets: *Human Activity Recognition* (HAR) dataset, d-dimension subspace: $d = 5, 10, 20, 50, 100$, and the dimension of the hypersphere is $n = 560$.

Data	dim	kmeans	Spkmeans	moVMF	LSC-HS	LSC-KL I	LSC-KL II	LSC-KL III
HAR	5	0.6001	0.5478	0.5579	0.5945	0.5645	**0.6236**	0.5994
	10				0.5970	0.6062	**0.6147**	0.5897
	20				0.5569	**0.6175**	0.6076	0.5849
	50				**0.6049**	0.5318	0.5726	0.5774
	100				**0.5875**	0.5278	0.5052	0.4849

$\hat{p}_M$ be the two marginal probability distributions. Let $I(p_J)$ be the mutual information and let H be entropy (Cover & Thomas 2006; MacKay 2005). The NMI is defined by

$$\text{NMI}(p_J) = I(p_J)/(H(p_M)H(\hat{p}_M))^{1/2}.$$

The values of the NMI are reported in Table 1, Table 2, and Table 3. Each value is an average over 10 runs. The entries highlighted in red are the best results for each dataset. Note that the Kmeans algorithm is used to initialize LSC-HS, and LSC-KL I, LSC-KL II, LSC-KL III are all initialized by LSC-HS.

In Appendix D the results obtained using the NMI are compared with the results using the Accuracy Rate and the Rand Index.

6.3 Experimental results

The NMI results of LSC-HS are consistently better than the results obtained from Kmeans and Spkmeans on different d-dimensional subspaces. Note that moVMF is a special case

Table 3: Comparison of NMI results for the datasets: *Gas Sensor Array Drift* (GSAD) datasets, d-dimension subspace: $d = 2, 4, 6, 8, 10, 15$, and the dimension of the hypersphere is $n = 127$.

Data	dim	kmeans	Spkmeans	moVMF	LSC-HS	LSC-KL I	LSC-KL II	LSC-KL III
GSAD	2	0.3311	0.3404	0.1346	0.3908	0.3101	0.3887	**0.3967**
	4				0.4050	0.3228	0.4396	**0.4658**
	6				0.4539	0.3693	0.4560	**0.4952**
	8				0.4727	0.3246	0.4583	**0.5259**
	10				0.4282	0.3003	0.4273	**0.5358**
	15				0.4572	0.2665	0.3944	**0.4902**

of LSC-HS: when $d = 1$, LSC-HS is equal to moVMF. The proposed LSC-HS is more flexible than moVMF as different d-dimension subspaces can be applied to the cluster problem.

Our proposed algorithms can be applied to different d-dimensional subspaces. The performance with higher values of d is no better than the performance with relatively low values. These phenomenons are illustrated in Table 1-Table 3, i.e., the best performance is obtained in the case of $d = 10$ for $\mathcal{D}_2$ and $\mathcal{D}_3$; $d = 20$ for subsets $\mathcal{D}_1$ and $\mathcal{D}_4$ in the mnist dataset; $d = 5$ in the HAR dataset; $d = 10$ in the GSAD dataset.

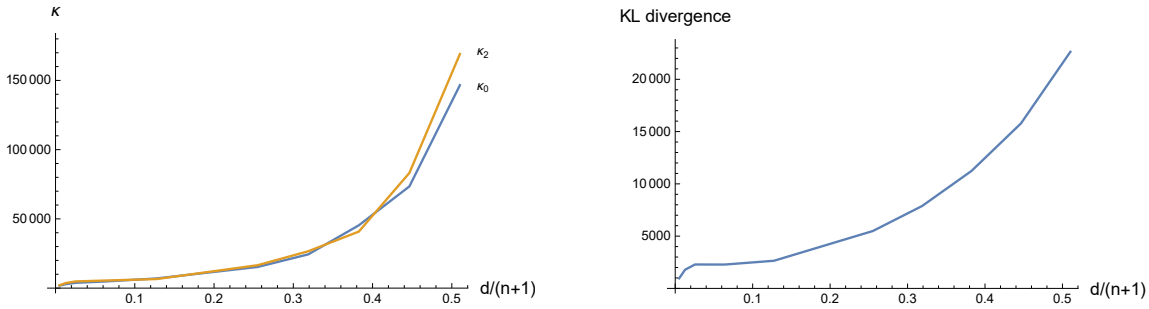
The performance of LSC-HS is better than that of LSC-KL in Table 1 and Table 3. The mnist data and the GSAD data yield better results for LSC-HS, compared with LSC-KL I. However, for the HAR dataset, the performance of LSC-KL I is usually better than that of LSC-HS as shown in Table 2.

The NMI results of LSC-KL II are better than those for LCS-KL I in Tables 1-3. This is due to the influence of the distance information on the KL divergence. If the distance between two clusters is larger than a threshold, then it is not necessary to obtain the KL divergence for the two clusters. However, a disadvantage of LSC-KL I and LSC-KL II is that once two clusters are merged it is not possible to separate them, if the merge turns out to be incorrect.

LSC-KL II and LSC-KL III each use LSC-HS for initialization. The key difference between LSC-KL II and LSC-KL III is that in LSC-KL III the initialization is carried out in each iteration. It is apparent from Table 1 and Table 3 that the NMI results of LSC-KL III are for the most part better than the results obtained for all other methods. However, for HAR, the NMI results of LSC-KL II are better than the results obtained for LSC-KL III as shown in Table 2.

6.4 Numerical examples

Some numerical results obtained by fitting pdfs in (2) to $\mathcal{C}_0$ and to $\mathcal{C}_2$ are shown in Fig. 2. Fig. 2a shows the concentration parameters κ_0 for $\mathcal{C}_0$ and κ_2 for $\mathcal{C}_2$ as functions of $d/(n+1)$. Fig. 2b shows the KL divergence from the pdf for $\mathcal{C}_2$ to the pdf for $\mathcal{C}_0$ as a function of $d/(n+1)$. The values of the KL divergence are high, which indicates that the two fitted pdfs are well separated.



(a) (b)

Figure 2: a) graphs of κ as a function of $d/(n+1)$ for pdfs fitted to $\mathcal{C}_0$ and $\mathcal{C}_2$; b) graph of the KL divergence of the pdf fitted to $\mathcal{C}_2$ from the pdf fitted to $\mathcal{C}_0$, as a function of $d/(n+1)$.

7 Conclusion

A new family of probability density functions for modelling clusters of vectors on a hypersphere is defined. Each pdf $p(x|P,\kappa)$, for x in the hypersphere, is conditional on an orthogonal projection P and a concentration parameter κ. The pdf takes its maximum value on all the points in the intersection of the range L of P with the hypersphere S^n. If κ is large, then $p(x|P,\kappa)$ is concentrated near to $L \cap S^n$. If κ is small, then $p(x|P,\kappa)$ approximates a uniform pdf on the hypersphere. The family of pdfs is parameterised by $G(d,n+1) \times \mathbb{R}_{\geq 0}$, where $G(d,n+1)$ is the Grassmann manifold for linear subspaces of dimension d in $\mathbb{R}^{n+1}$.

The pdfs provide a wide range of models for clusters, while at the same time the pdfs are tractable in that *i)* a given pdf can be sampled efficiently; *ii)* the parameters of a pdf can be estimated using maximum likelihood and *iii)* the Kullback-Leibler divergence and the Fisher-Rao metric can be evaluated with a low time complexity. The KL divergence is used to compare clusters of vectors. If the KL divergence is small then the two clusters in question are similar, and thus candidates for merging. The values of the KL divergence are independent of the choice of parameterisation of the manifold $G(d,n+1) \times \mathbb{R}_{\geq 0}$. Thus the KL divergence is not affected by accidental properties of the parameterisation that have nothing to do with clustering.

If d and n are large, then numerical experiments indicate that there is a background value for the KL divergence. If the concentration κ is fixed and if P_1, P_2 are projections sampled independently from the uniform distribution on $G(d,n+1)$ defined by the Fisher-Rao metric, then the KL divergence of (P_2,κ) from (P_1,κ) is closely approximated by an explicit function of d, n, and κ. This observation relies on the fact that $G(d,n+1)$ is compact. There is no corresponding result for Gaussian distributions on $\mathbb{R}^{n+1}$. In this context, a value of the KL divergence is small if it is significantly less than the background value.

The dimension d of the range of a projection P in $G(d,n+1)$ can be varied in order to obtain the best fit pdf to a cluster. In high dimensions, the best fit value of d is unlikely to be small. This is because any two random vectors u, v in S^n for n large are likely to be near orthogonal in that the scalar product $u.v$ is small (Gorban & Tyukin 2018).

Four algorithms that use the family of pdfs for clustering vectors on a hypersphere are described. Three of the algorithms use a symmetrized version of the Kullback-Leibler divergence to compare the different clusters. The algorithms are tested on three datasets, namely mnist, Human Activity Recognition and Gas Sensor Array Drift with good results. The best performing algorithm, LSC-KL III, uses the KL divergence iteratively to merge pairs of clusters, in order to reduce the number of clusters from a starting value of c_0 to a required value c, $c \leq c_0$.

The accuracy of the clustering is assessed in Section 6 using the Normalised Mutual Information (NMI). In Appendix D the NMI is replaced by the Accuracy Rate (AR) and the Rand Index (RI). The two algorithms, LSC-HS and LSC-KL III, that perform well using NMI also perform well using AR and RI.

References

1. M. Abramowitz & I.A. Stegun (1965) *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover Publications Inc., New York.

2. S.-I. Amari (1985) *Differential-Geometric Methods in Statistics*. Lecture Notes in Statistics, vol. 28. Springer-Verlag.

3. D. Anguita, A. Ghio, L. Oneto, X. Parra & J. L. Reyes-Ortiz (2013) A public domain dataset for human activity recognition using smartphones. *21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, ESANN 2013, Bruges, Belgium, 24-26 April 2013.

4. A. Banerjee, I. S. Dhillon, J. Ghosh, & S. Sra. (2005)a Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, vol. 6(Sep): 13451382.

5. A. Banerjee, S. Merugu, I.S. Dhillon & J. Ghosh (2005)b Clustering with Bregman divergencies. *Journal of Machine Learning Research*, vol. 6(Oct): 1705-1749.

6. C. Bingham (1974) An antipodally symmetric distribution on the sphere. *Annals of Statistics*, vol. 2, no. 6, pp. 1201-1225.

7. Y. Chikuse (2003) *Statistics on Special Manifolds*. Lecture Notes in Statistics, vol. 174, Springer.

8. Cover, T.M. & Thomas, J.A. (2006) *Elements of Information Theory.* 2nd Edition, Wiley-Interscience.

9. I.S. Dhillon & D.S. Modha (2001) Concept decompositions for large sparse text data using clustering. *Machine Learning*, vol. 42(1-2): 143-175.

10. E. Garcia-Portugués, D. Paindaveine & T. Verdebout (2020) On optimal tests for rotational symmetry against new classes of hyperspherical distributions. *Journal of the American Statistical Association*, vol. 115, pp. 1873-1887.

11. A. N. Gorban & I. Y. Tyukin (2018) Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A, mathematical, physical and engineering sciences*, vol. 376.

12. G.B. Folland (2001) How to integrate a polynomial over a sphere. *American Mathematical Monthly*, vol. 108, no. 5, pp. 446-448.

13. O.C. Hamsici & A.M. Martinez (2007) Spherical-homoscedastic distributions: the equivalency of spherical and normal distributions in classification. *Journal of Machine Learning Research*, vol. 8, pp. 1583-1623.

14. A.T. James (1954) Normal multivariate analysis and the orthogonal group. *Annals of Mathematical Statistics*, vol. 25, no. 1, pp. 40-75.

15. J.T. Kent, A.M. Ganeiber & K.V. Mardia (2018) A new unified approach for the simulation of a wide class of directional distributions. *Journal of Computational and Graphical Statistics*, vol. 27, issue 2, pp. 291-301.

16. G. Kurz, F. Pfaff & U.D. Hanebeck (2016) Kullback-Leibler divergence and moment matching for hyperspherical probability distributions. *Proceedings of the 19th International Conference on Information Fusion (Fusion 2016)*, Heidelberg, Germany.

17. Y. LeCun, L Bottou, Y. Bengio & P. Haffner (1998) Gradient based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324.

18. E. Ley & T. Verdebout (2017) *Modern Directional Statistics*. Chapman and Hall/CRC.

19. E. Ley & T. Verdebout (eds.) (2018) *Applied Directional Statistics: modern methods and case studies*. Chapman and Hall/CRC Interdisciplinary Statistics, 1st edition.

20. D. MacKay (2005) *Information theory, Inference, and Learning*. Cambridge University Press.

21. K.V. Mardia (1975) Statistics of directional data. *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 37, no. 3, pp. 349-393.

22. K.V. Mardia & P.E. Jupp (1999) *Directional Statistics*. John Wiley and Sons Ltd.

23. P.J. Paine, S.P. Preston, M. Tsagris & A.T.A. Wood (2017) An elliptically symmetric angular Gaussian distribution. *Statistics and Computing*, vol. 28, no. 3, pp. 689-697.

24. A. Painsky & G.W. Wornell (2019) Bregman divergence bounds and universality properties of the logarithmic loss. *IEEE Transactions on Information Theory*, Vol 66(3), pp. 1658-1673.

25. A. Pewsey & E. Garcia-Portugués (2021) Recent advances in directional statistics. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*. Springer, vol. 30, no. 1, pp. 1-58.

26. J.G. Saw (1978) A family of distributions on the m-sphere and some hypothesis tests. *Biometrika*, Vol 65, No 1, pp 69-73.

27. J.L. Scealy & A.T.A. Wood (2019) Scaled von Mises-Fisher distributions and regression models for paleomagnetic directional data. *Journal of the American Statistical Association*, vol. 114, issue 528, pp. 1547-1560.

28. B. Selby (1964) Girdle distributions on a sphere. *Biometrika*, vol. 51, pp. 381-392.

29. S. Sra. (2018) *Directional statistics in machine learning: a brief review.* In E. Ley & T. Verdebout (eds.) *Applied Directional Statistics: modern methods and case studies*, ch. 12, pp. 259-276. Chapman and Hall/CRC Interdisciplinary Statistics, 1st edition.

30. S. Sra and D. Karp (2013) The multivariate Watson distribution: maximum likelihood estimation and other aspects. *Journal of Multivariate Analysis*, Vol 114, pp. 256-269.

31. A. Strel & J. Ghosh (2002) Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, vol. 3(Dec): 583-617.

32. A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer & R. Huerta (2012) Chemical gas sensor drift compensation using classifier ensembles, *Sensors and Actuators B: Chemical*, doi: 10.1016/j.snb.2012.01.074.

33. N.X. Vinh & J. Epps (2009) A novel approach for automatic number of clusters detection in microarray data based on consensus clustering. In *Proceedings 9th IEEE International Conference on Bioinformatics and Bioengineering (Taichung, Taiwan)*, pp. 84-91.

34. N.X. Vinh, J. Epps & J. Bailey (2009) Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings 26th International Conference on Machine Learning (ICML09), ACM*, pp. 1073-1080.

35. N.X. Vinh, J. Epps & J. Bailey (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, vol. 11(Oct): 2837-2854.

36. F. Wang, X. Xiang, J. Cheng & A. L. Yuille (2017) NormFace: L_2 hypersphere embedding for face verification. *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*, pp. 1041-1049.

37. J. Yang, E. Grunsky & Q. Cheng (2019) A novel hierarchical clustering analysis method based on Kullback-Leibler divergence and application on dalaimiao geochemical exploration data. *Computers and Geosciences*, vol. 123, pp. 10-19.

38. J. Zhang, G. Zhu, R.W. Heath & K. Huang (2018) Grassmannian learning; embedding geometry awareness in shallow and deep learning. *arXiv:1808.02229.*

39 C. Zhao & J.S. Song (2018) Exact heat equation on a hypersphere and its applications in kernel SVM. *Frontiers in Applied Mathematics and Statistics*, vol. 1, article 1. DOI 10.3389/fams.2018.00001

A Integration over S^n

The terms $\langle x^\top E(I-Q)x\rangle$ and $\langle \|Ex\|^2\rangle$ on the right hand side of (18) are simplified. The subscript 1 in (18) is omitted here. Recall from Section 4.2 that x, y, z are vectors such that y is in $\mathbb{R}^d$, z is in $\mathbb{R}^{n-d+1}$ and $x = (y^\top, z^\top)^\top$. Let Q be the $(n+1)\times(n+1)$ projection matrix defined by (11). It follows that $\|(I-Q)x\|^2 = \|z\|^2$, thus the pdf (2) for the generalised Watson distribution reduces to

$$p(x|Q,\kappa) = C_n(\kappa)\exp\left(-\kappa\|z\|^2/2\right), \quad x \in S^n. \tag{31}$$

It is convenient to use $\langle . \rangle$ for integration over S^n with the weight $p(x|Q,\kappa)$. The invariance of $p(x|Q,\kappa)$ under the action of the $(n+1)\times(n+1)$ orthogonal matrices on $G(d,n+1)$ ensures that

$$\begin{aligned} \langle y_i\rangle &= 0, \qquad 1 \le i \le d, \\ \langle y_i y_j\rangle &= 0, \qquad 1 \le i,j \le d, i \ne j. \\ \langle y_i^2\rangle &= \langle y_j^2\rangle, \quad 1 \le i,j \le d. \end{aligned}$$

Similar results hold for $\langle z_i\rangle$ and $\langle z_i z_j\rangle$. In addition, $\langle y_i z_j\rangle = 0$ for $1 \le i \le d$ and $1 \le j \le n-d+1$. See Folland (2001).

Let the $(n+1)\times(n+1)$ symmetric matrix E in the expression (18) for the KL divergence have the block structure shown in (17). The matrix E is chosen such that $Q+E$ is a projection matrix. Let tr be the trace function for matrices. It is noted that the projection matrices corresponding to points in $G(d,n+1)$ all have trace d. In particular,

$$d = \mathrm{tr}(Q) = \mathrm{tr}(Q+E), \tag{32}$$

which yields $\mathrm{tr}(E) = 0$. It follows from (17) that $\mathrm{tr}(A) = \mathrm{tr}(B)$. A short calculation yields

$$\begin{aligned} \left\langle (y^\top, z^\top)E(I-Q)\begin{pmatrix} y \\ z\end{pmatrix}\right\rangle &= \left\langle (y^\top, z^\top)E\begin{pmatrix} 0 \\ z\end{pmatrix}\right\rangle, \\ &= \langle z^\top B z\rangle, \\ &= \mathrm{tr}(B)(n-d+1)^{-1}\langle \|z\|^2\rangle, \\ &= \mathrm{tr}(A)(n-d+1)^{-1}\langle \|z\|^2\rangle. \end{aligned} \tag{33}$$

Next, the expression $\langle \|Ex\|^2\rangle$ is simplified. The matrix $Q+E$ is a projection matrix, thus

$$(Q+E)(Q+E) = Q+E,$$

which reduces to

$$\begin{pmatrix} -A & 0 \\ 0 & 0\end{pmatrix} + EE = \begin{pmatrix} 0 & 0 \\ 0 & B\end{pmatrix}. \tag{34}$$

It follows from (34) that

$$\begin{aligned}\left\langle \left\| E\begin{pmatrix} y \\ z \end{pmatrix} \right\|^2 \right\rangle &= \left\langle (y^\top, z^\top) EE \begin{pmatrix} y \\ z \end{pmatrix} \right\rangle, \\ &= \langle y^\top A y \rangle + \langle z^\top B z \rangle, \\ &= \mathrm{tr}(A) d^{-1} \langle \|y\|^2 \rangle + \mathrm{tr}(B)(n-d+1)^{-1} \langle \|z\|^2 \rangle, \\ &= \mathrm{tr}(A)(d^{-1}\langle \|y\|^2 \rangle + (n-d+1)^{-1}\langle \|z\|^2 \rangle). \end{aligned} \tag{35}$$

B Evaluation of Integrals

As in Section 4.2, let x, y, z be vectors such that x is in S^n, y is in $\mathbb{R}^d$, z is in $\mathbb{R}^{n-d+1}$ and $x = (y^\top, z^\top)^\top$. In this Appendix expressions are obtained for $C_n(\kappa)$ (Sections 3.2 and 3.3), $\langle \|y\|^2 \rangle$ (Section 4.2) and $\langle \|z\|^2 \rangle$ (Section 4.2). The expressions involve confluent hypergeometric functions (Abramowitz & Stegun 1965).

Let i be a non-negative integer, and let K_i be the integral

$$K_i = \int_{S^n} \|z\|^{2i} \exp(-(\kappa/2)\|z\|^2)\, d\omega_n, \tag{36}$$

where $d\omega_n$ is the measure induced on S^n by the Lebesgue measure in $\mathbb{R}^{n+1}$. It follows that

$$\begin{aligned} C_n(\kappa)^{-1} &= K_0, \\ \langle \|y\|^2 \rangle &= 1 - \langle \|z\|^2 \rangle, \\ \langle \|z\|^2 \rangle &= K_1/K_0. \end{aligned} \tag{37}$$

Expressions for the K_i are obtained. Let γ_i be the volume of the i-dimensional hypersphere with unit radius,

$$\gamma_i = 2\pi^{(i+1)/2}/\Gamma((i+1)/2),$$

where Γ is the Gamma function (Abramowitz & Stegun 1965). It is noted that

$$\int_0^\infty r^s \exp(-r^2)\, dr = 2^{-1}\Gamma((s+1)/2).$$

Let i be a non-negative integer and let H_i be defined by

$$H_i = \int_{S^n} \|z\|^{2i}\, d\omega_n. \tag{38}$$

An alternative expression for H_i is obtained. With this in mind, let F_i be defined by

$$F_i = \int_{\mathbb{R}^{n+1}} \|z\|^{2i} \exp\left(-\|y\|^2 - \|z\|^2\right)\, dydz. \tag{39}$$

The integral over $\mathbb{R}^{n+1}$ on the right hand side of (39) is evaluated in two different ways. In the first way, the integral is split into two independent integrals and each of these

integrals is reduced to a one dimensional integral using polar coordinates,

$$\begin{aligned} F_i &= \left(\int_{\mathbb{R}^d} \exp(-\|y\|^2)\, dy\right)\left(\int_{\mathbb{R}^{n-d+1}} \|z\|^{2i} \exp(-\|z\|^2)\, dz\right), \\ &= \left(\gamma_{d-1}\int_0^\infty r^{d-1}\exp(-r^2)\, dr\right)\left(\gamma_{n-d}\int_0^\infty r^{2i+n-d}\exp(-r^2)\, dr\right), \\ &= 4^{-1}\gamma_{d-1}\gamma_{n-d}\Gamma(d/2)\Gamma(i+(n-d+1)/2). \end{aligned} \tag{40}$$

In the second way, the integral in (39) is reduced to a one dimensional integral by taking polar coordinates in $\mathbb{R}^{n+1}$,

$$\begin{aligned} F_i &= \int_{S^n}\int_0^\infty \|z\|^{2i} r^n \exp(-r^2)\, dr d\omega_n, \\ &= \int_{S^n}\int_0^\infty \|r^{-1}z\|^{2i} r^{n+2i}\exp(-r^2)\, dr d\omega_n, \\ &= \left(\int_{S^n}\|z\|^{2i}\, d\omega_n\right) 2^{-1}\Gamma(i+(n+1)/2), \\ &= H_i 2^{-1}\Gamma(i+(n+1)/2). \end{aligned} \tag{41}$$

It follows from (40) and (41) that

$$H_i = 2^{-1}\gamma_{d-1}\gamma_{n-d}\Gamma(i+(n+1)/2)^{-1}\Gamma(d/2)\Gamma(i+(n-d+1)/2). \tag{42}$$

The result (42) is used to obtain an alternative expression for K_i. The exponential function in (36) is expanded to yield

$$\begin{aligned} K_i &= \int_{S^n}\|z\|^{2i}\left(\sum_{k=0}^\infty \frac{\|z\|^{2k}(-\kappa/2)^k}{k!}\right) d\omega_n, \\ &= \sum_{j=0}^\infty H_{i+j}\frac{(-\kappa/2)^j}{j!}, \\ &= 2^{-1}\gamma_{d-1}\gamma_{n-d}\Gamma(d/2)\sum_{j=0}^\infty \frac{\Gamma(i+j+(n-d+1)/2)(-\kappa/2)^j}{j!\Gamma(i+j+(n+1)/2)}. \end{aligned} \tag{43}$$

Let $(q)_k$ be the Pochhammer symbol defined by

$$\begin{aligned} (q)_0 &= 1, \\ (q)_k &= q(q+1)\ldots(q+k-1), \quad k=1,2,3,\ldots, \\ &= \Gamma(q)^{-1}\Gamma(q+k). \end{aligned}$$

The confluent hypergeometric function ${}_1F_1(a,b,z)$ is defined in Section 13.1.2 of Abramowitz & Stegun (1965) by

$${}_1F_1(a,b,z) = \sum_{k=0}^\infty \frac{(a)_k z^k}{k!(b)_k}, \tag{44}$$

Abramowitz & Stegun (1965) use the notation $M(a,b,z)$ instead of ${}_1F_1(a,b,z)$, and refer to $M(a,b,z)$ as Kummer's function. The regularized confluent hypergeometric function ${}_1\tilde{F}_1(a,b,z)$ is defined by

$$ {}_1\tilde{F}_1(a,b,z) = {}_1F_1(a,b,z)/\Gamma(b). \tag{45}$$

It follows from (43),(44) and (45) that

$$K_i =$$

$$2^{-1}\gamma_{d-1}\gamma_{n-d}\Gamma(d/2)\Gamma(i+(n-d+1)/2){}_1\tilde{F}_1(i+(n-d+1)/2, i+(n+1)/2, -\kappa/2). \tag{46}$$

The scale factor $C_n(\kappa)$ in Section 3.3 is given by

$$\begin{aligned} C_n(\kappa)^{-1} &= K_0, \\ &= 2^{-1}\gamma_{d-1}\gamma_{n-d}\Gamma(d/2)\Gamma((n-d+1)/2)\ {}_1\tilde{F}_1((n-d+1)/2,(n+1)/2,-\kappa/2), \\ &= 2\pi^{(n+1)/2}{}_1\tilde{F}_1((n-d+1)/2,(n+1)/2,-\kappa/2). \end{aligned} \tag{47}$$

It follows from (37) that

$$\langle \|z\|^2 \rangle = \frac{(n-d+1)\,{}_1\tilde{F}_1((n-d+3)/2,(n+3)/2,-\kappa/2)}{2\ {}_1\tilde{F}_1((n-d+1)/2,(n+1)/2,-\kappa/2)}. \tag{48}$$

The value of $\langle \|y\|^2 \rangle$ is obtained by observing that

$$\langle \|y\|^2 \rangle = 1 - \langle \|z\|^2 \rangle = 1 - K_1/K_0. \tag{49}$$

It follows from (36) that

$$\frac{\partial K_0}{\partial \kappa} = -2^{-1}K_0\langle \|z\|^2 \rangle,$$

thus

$$\langle \|z\|^2 \rangle = 2\frac{\partial}{\partial \kappa}\ln(C_n(\kappa)). \tag{50}$$

C Approximations

The terms $\langle \|z\|^2 \rangle$ (Section 4.2), g (Section 4.2) and f (Section 4.4) are approximated for large values of the concentration parameter κ.

The expression (48) for $\langle \|z\|^2 \rangle$ contains a ratio of regularised confluent hypergeometric functions ${}_1\tilde{F}_1$, as defined by (45). Let ι be the square root of -1. The following approximation to ${}_1\tilde{F}_1$ is obtained from Section 13.5.1 of Abramowitz & Stegun (1965), with minor changes in notation,

$$ {}_1\tilde{F}_1(a,b,-\kappa/2) =$$

$$\begin{aligned} &= \Gamma(b-a)^{-1}\exp(\pm\iota\pi a)(-\kappa/2)^{-a}\left(\sum_{n=0}^{R-1}\frac{2^n(a)_n(1+a-b)_n}{n!\kappa^n} + O(\kappa^{-R})\right) \\ &\quad + \Gamma(a)^{-1}\exp(-\kappa/2)(-\kappa/2)^{a-b}\left(\sum_{n=0}^{S-1}\frac{(b-a)_n(1-a)_n(-2)^n}{n!\kappa^n} + O(\kappa^{-S})\right), \end{aligned} \tag{51}$$

where $(.)_n$ is the Pochhammer symbol defined in Appendix B. The upper sign is taken in (51) if $-\pi/2 < \arg(-\kappa/2) < 3\pi/2$ and the lower sign is taken if $-3\pi/2 < \arg(-\kappa/2) < -\pi/2$. It is convenient to set the value of $\arg(-\kappa/2)$ equal to π. It follows that

$$\exp(\iota\pi a)(-\kappa/2)^{-a} = \exp(\iota\pi a)(\kappa/2)^{-a}\exp(-\iota\pi a) = (\kappa/2)^{-a}.$$

The second summation in (51) has a factor $\exp(-\kappa/2)$ which is negligible if κ is large. The index R is set equal to 2 in the first summation. It follows that

$${}_1\tilde{F}_1(a, b, -\kappa/2) = \Gamma(b-a)^{-1}(\kappa/2)^{-a}\left(1 + 2a(1+a-b)\kappa^{-1} + O(\kappa^{-2})\right). \tag{52}$$

It follows from (48) and (52) that

$$\langle \|z\|^2 \rangle = \kappa^{-1}(n-d+1)\left(1 + (2-d)\kappa^{-1} + O(\kappa^{-2})\right) \tag{53}$$

An approximation is obtained for the KL divergence (20) with $\kappa = \kappa_1 = \kappa_2$. It follows from (20) that

$$D(Q, \kappa \| Q + E, \kappa) = 2^{-1}\mathrm{tr}(A)\kappa g(d, n, \kappa) \tag{54}$$

where

$$\begin{aligned} g(d, n, \kappa) &\equiv d^{-1}\langle \|y\|^2 \rangle - (n-d+1)^{-1}\langle \|z\|^2 \rangle, \\ &= d^{-1} - d^{-1}\langle \|z\|^2 \rangle - (n-d+1)^{-1}\langle \|z\|^2 \rangle, \\ &= d^{-1}(1 - \kappa^{-1}(n+1)) + O(\kappa^{-2}). \end{aligned} \tag{55}$$

It follows from (54) and (55) that

$$D(Q, \kappa \| Q + E, \kappa) = 2^{-1}\mathrm{tr}(A)d^{-1}(\kappa - n - 1) + O(\kappa^{-1}).$$

The normalising factor in the generalised Watson distribution (2) is given by

$$\begin{aligned} C_n(\kappa)^{-1} &= K_0, \\ &= 2\pi^{(n+1)/2}{}_1\tilde{F}_1((n-d+1)/2, (n+1)/2, -\kappa/2), \end{aligned}$$

$$= 2^{(n-d+3)/2}\pi^{(n+1)/2}(\Gamma(d/2)\kappa^{(n-d+1)/2})^{-1}\left(1 + (n-d+1)(1-d/2)\kappa^{-1} + O(\kappa^{-2})\right).$$

The expression (27) for the Fisher-Rao metric includes the function f defined by (23). It follows from (50), (53) and the definition of f that

$$\begin{aligned} f(\kappa) &= -2^{-1}\frac{\partial}{\partial\kappa}\langle \|z\|^2 \rangle, \\ &= 2^{-1}\kappa^{-2}(n-d+1) + O(\kappa^{-3}). \end{aligned} \tag{56}$$

D Clustering Evaluation

In this appendix two methods, namely *accuracy rate* (AR) and *rand index* (RI), are used to measure the statistical similarity between a cluster and the ground truth. The two methods are tested on the three datasets mnist, Human Activity Recognition, and Gas Sensor Array Drift. The results are compared with those obtained in Section 6 using the Normalised Mutual Information. The parameter settings are the same as for the NMI in Section 6.

D.1 Accuracy rate

Given N samples in a dataset, let y_i be the class label for the i-th sample and let $\hat{y}_i$ be the predicted class label. The accuracy rate (AR) between y and $\hat{y}$ is defined by finding the best match between the class labels and the cluster labels:

$$\mathrm{AR}(y, \hat{y}) = \max_{\mathrm{perm} \in \mathcal{P}} \frac{1}{N} \sum_{i=0}^{N-1} \mathbb{I}(\mathrm{perm}(\hat{y}_i) = y_i), \tag{57}$$

where $\mathcal{P}$ is the set of all permutations in $\{1, ..., K\}$, K is the number of clusters, and $\mathbb{I}(\cdot)$ is the indicator function (i.e., $\mathbb{I}(\hat{y}_i = y_i) = 1$ if $\hat{y}_i = y_i$ and 0 otherwise).

Table 4, Table 5, and Table 6 show the values of AR for the algorithms. It can be seen that 1) The AR results of LSC-HS are better than those of Kmeans, Spkmeans and moVMF on the different d-dimensional subspaces. This is consistent with the performance of the algorithms as measured by NMI. 2) For the datasets GSAD, and $\mathcal{D}_2$-$\mathcal{D}_4$, the AR performance of LSC-HS is better than that of LSC-KL III. For the datasets HAR, and $\mathcal{D}_1$, the AR performance of LSC-KL III is better than that of LSC-HS. 3) The performances using the low-dimensional subspace are better than those for the high-dimensional subspaces, i.e., the best performance is obtained under $d = 5$ for HAR dataset, $d = 6$ for GSAD, $d = 10$ for $\mathcal{D}_3$, and $d = 20$ for $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_4$.

Table 4: Comparison of *Accuracy Rate* (%) results for the datasets: $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$. The dimension of the hypersphere is $n = 783$ and the subspace dimension $d = 10, 20, 50, 100$.

Data	dim	kmeans	Spkmeans	moVMF	LSC-HS	LSC-KL I	LSC-KL II	LSC-KL III
$\mathcal{D}_1$	10	96.68	96.33	96.07	99.06	76.31	96.01	**99.20**
	20				99.01	76.81	98.95	**99.25**
	50				**99.06**	74.91	94.73	98.54
	100				97.20	71.55	92.67	**97.63**
$\mathcal{D}_2$	10	75.47	75.49	74.83	**79.55**	58.40	65.12	58.16
	20				**80.13**	53.17	71.60	61.06
	50				**78.18**	51.18	70.76	67.14
	100				**76.00**	52.58	65.25	66.10
$\mathcal{D}_3$	10	57.87	57.04	52.36	**60.91**	47.66	45.04	54.60
	20				**57.28**	42.92	47.04	50.44
	50				**56.71**	36.08	45.86	41.66
	100				**58.28**	39.21	45.17	39.59
$\mathcal{D}_4$	10	53.98	54.91	20.11	**58.63**	30.68	32.27	44.96
	20				**59.37**	32.52	34.20	45.71
	50				**58.23**	29.57	32.35	36.41
	100				**56.35**	28.66	34.45	35.35

Table 5: Comparison of *Accuracy Rate* (%) results for the Human Activity Recognition (HAR) datasets. The dimension of the hypersphere is $n = 560$, and the subspace dimension $d = 5, 10, 20, 50, 100$.

Data	dim	kmeans	Spkmeans	moVMF	LSC-HS	LSC-KL I	LSC-KL II	LSC-KL III
HAR	5	58.07	51.86	58.38	57.39	45.64	53.92	**58.41**
	10				**56.16**	47.74	50.11	55.44
	20				53.21	48.76	51.55	**55.60**
	50				**58.12**	45.33	51.28	54.94
	100				**57.29**	44.62	41.85	41.55

Table 6: Comparison of *Accuracy Rate* (%) results for the Gas Sensor Array Drift (GSAD) datasets. The dimension of the hypersphere is $n = 127$ and the subspace dimension $d = 2, 4, 6, 8, 10, 15$.

Data	dim	kmeans	Spkmeans	moVMF	LSC-HS	LSC-KL I	LSC-KL II	LSC-KL III
GSAD	2	43.26	43.70	31.90	52.00	41.21	45.26	**53.55**
	4				50.32	41.16	46.34	**52.53**
	6				55.96	43.39	48.30	**59.09**
	8				54.37	39.01	47.74	**55.32**
	10				52.45	37.42	46.56	**54.16**
	15				**57.68**	34.54	42.06	50.72

D.2 Rand index

The rand index (RI) is a way of comparing the results obtained by two different clustering methods. The formula for RI is

$$\mathrm{RI} = \binom{N}{2}^{-1}(N_{11} + N_{00}),$$

where N_{11} is the number of times a pair of elements belongs to the same cluster across two clustering methods, N_{00} is the number of times a pair of elements belong to different clusters across two clustering methods, and $\binom{N}{2}$ is the number of unordered pairs in a set of N elements.

Table 7, Table 8, and Table 9 show the RI performance of the algorithms. It can be seen that 1) For the mnist dataset, the AR results of LSC-HS are better than those of Kmeans, Spkmeans, moVMF for the different d-dimensional subspaces. This is consistent with the performance of NMI and AR. On the subset $\mathcal{D}_1$, the RI performance of LSC-KL III is better than that of LSC-HS. For other subsets $\mathcal{D}_2$-$\mathcal{D}_4$, LSC has the best RI results. 2) For the HAR dataset, LSC-KL III has the best RI results compared with the other algorithms. 3) For the GSAD dataset, LSC has the best RI results for the different d-dimensional subspaces.

Table 7: Comparison of *Rand Index* (%) results for the datasets: $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$. The dimension of the hypersphere is $n = 783$ and the subspaces have dimensions $d = 10, 20, 50, 100$.

Data	dim	kmeans	Spkmeans	moVMF	LSC-HS	LSC-KL I	LSC-KL II	LSC-KL III
$\mathcal{D}_1$	10	93.58	92.92	92.45	98.13	64.87	93.90	**98.41**
	20				98.03	65.37	97.92	**98.52**
	50				**98.13**	63.74	91.70	97.19
	100				94.56	61.12	88.36	**95.38**
$\mathcal{D}_2$	10	76.45	76.49	76.79	**80.67**	73.03	74.11	73.42
	20				**80.92**	67.39	77.51	73.22
	50				**78.97**	63.72	76.05	74.84
	100				**76.90**	58.39	71.85	74.26
$\mathcal{D}_3$	10	79.75	79.58	72.80	**81.05**	72.90	68.87	80.23
	20				**79.94**	66.04	73.46	79.41
	50				**79.58**	57.76	69.51	67.04
	100				**79.84**	60.28	67.50	62.99
$\mathcal{D}_4$	10	88.26	88.51	53.18	**89.47**	63.74	66.79	86.01
	20				**89.54**	63.47	68.61	83.39
	50				**89.07**	62.36	70.93	78.31
	100				**88.69**	62.55	70.41	71.67

Table 8: Comparison of *Rand Index* (%) results for the Human Activity Recognition (HAR) dataset. The dimension of the hypersphere is $n = 560$, and the subspaces have dimensions $d = 5, 10, 20, 50, 100$.

Data	dim	kmeans	Spkmeans	moVMF	LSC-HS	LSC-KL I	LSC-KL II	LSC-KL III
HAR	5	82.84	76.97	83.59	82.85	76.50	81.01	**83.99**
	10				83.04	79.28	78.92	**83.49**
	20				80.22	78.91	77.96	**81.94**
	50				**83.00**	73.14	76.60	80.76
	100				**82.81**	73.43	71.47	72.70

Table 9: Comparison of *Rand index* (%) results for the Gas Sensor Array Drift (GSAD) datasets. The dimension of the hypersphere is $n = 127$ and the subspaces have dimensions $d = 2, 4, 6, 8, 10, 15$.

Data	dim	kmeans	Spkmeans	moVMF	LSC-HS	LSC-KL I	LSC-KL II	LSC-KL III
GSAD	2	68.40	64.73	64.7315	**78.31**	61.73	69.75	76.87
	4				**79.28**	62.01	72.13	77.39
	6				**80.51**	66.18	75.50	80.07
	8				**79.85**	63.50	70.11	79.42
	10				**79.07**	62.40	68.83	78.80
	15				**79.45**	59.93	64.79	73.73