A solution for the mean parametrization of the von Mises-Fisher distribution

Anonymous Authors[1]

Abstract

The von Mises-Fisher distribution as an exponential family can be expressed in terms of either its natural or its mean parameters. Unfortunately, however, the normalization function for the distribution in terms of its mean parameters is not available in closed form, limiting the practicality of the mean parametrization and complicating maximum-likelihood estimation more generally. We derive a second-order ordinary differential equation, the solution to which yields the mean-parameter normalizer along with its first two derivatives, as well as the variance function of the family. We also provide closed-form approximations to the solution of the differential equation. This allows rapid evaluation of both densities and natural parameters in terms of mean parameters, which are estimated most easily in the course of learning tasks involving directional statistics. We show applications to topic modeling with mixtures of von Mises-Fisher distributions using Bregman Clustering.

1. Introduction

The von Mises-Fisher (vMF) distribution is a classical distribution for data located on the unit hypersphere. It is an important tool in directional statistics (Mardia, 1975) and emerges naturally when considering data in the form of normalised vectors or embeddings (Banerjee et al., 2005a) as is common in machine learning. It is the maximum entropy distribution given the first moment $\mathbb{E}[X]$ of a variable defined on the unit hypersphere $X \in S^{D-1}$ (Rao, 1973), and has been used for clustering of angular data (Banerjee et al., 2005a), and more recently as a latent variable distribution in a variational autoencoder framework (Davidson et al., 2018; Xu & Durrett, 2018).

In exponential family form, the vMF distribution can be described by the density

$$p(x|\eta) = \frac{1}{(2\pi)^{D/2}}\, e^{\eta^\top x - \Phi(\eta)},$$

i.e. with base measure $\chi(x) = (2\pi)^{-D/2}$ uniform over $x \in S^{D-1}$, natural parameter $\eta \in \mathbb{R}^D$ and log-partition function

$$\begin{aligned}\Phi(\eta) &= \log \int e^{\eta^\top x} d\chi \\ &= \log I_{D/2-1}(||\eta||) - (D/2-1)\log||\eta||,\end{aligned}$$

where $||\cdot||$ is the Euclidean norm and I_v denotes the modified Bessel function of the first kind and order v. As a natural exponential family distribution with strictly convex log-partition function, a vMF distribution can equivalently be parametrized by the expectation $\mu = \mathbb{E}[X] = \nabla\Phi(\eta)$, also known as the mean parameter[1], in the form

$$p(x|\mu) = \frac{1}{(2\pi)^{D/2}}\, e^{\nabla\Psi(\mu)^\top(x-\mu)+\Psi(\mu)},$$

where

$$\Psi(\mu) = \Phi^*(\mu) = \max_\eta \mu^\top\eta - \Phi(\eta)$$

is the Legendre transform $\Psi = \Phi^*$ of the log-partition function Φ. Because $\Psi(\mu) = -H[X|\mu] - \mathbb{E}[\log\chi(X)|\mu]$ and because $\chi(x)$ is uniform for the vMF distribution, we will refer to Ψ as the "negative entropy" function. Ψ is strictly convex and of Legendre-type (Rockafellar, 1997), i.e. $\lim_{||\mu||\to 1}||\nabla\Psi(\mu)|| = \infty$ as μ approaches the domain boundary of Ψ at $||\mu|| = 1$.

The mean-parametrized form has several nice properties such as a well-interpretable domain for Ψ (Wainwright & Jordan, 2008), and is advantageous for learning as can be seen with a simple example: assume we have independently generated angular data $x^n \in \mathbb{S}^{D-1}, n = 1,\dots,N$, and want to fit a vMF distribution to it. In mean parametrization, the log-likelihood is given by

$$\begin{aligned}\mathrm{LL}(\mu) &= \frac{1}{N}\sum_{n=1}^N \log p(x^n|\mu) \\ &= -D_\psi\left(\frac{1}{N}\sum_{n=1}^N x^n \,||\, \mu\right) + \frac{1}{N}\sum_{n=1}^N \Psi(x^n) + \log\chi(x^n)\end{aligned}$$

[1] A common notation for the von Mises-Fisher parameters is $(\mu,\kappa) \in S^{D-1}\times\mathbb{R}_{\geq 0}$ with direction μ and concentration parameter κ. Here, we reserve μ with $||\mu|| \leq 1$ to refer to the mean parameter in exponential family form, giving direction and concentration parameters of $\eta/||\eta|| = \mu/||\mu||$ and $\kappa = ||\eta|| = ||\nabla\Psi(\mu)||$, respectively.

with non-negative Bregman divergence

$$D_\psi(x||\mu) = \nabla\Psi(\mu)^\top(\mu - x) - \Psi(\mu) + \Psi(x) \geq 0.$$

Because $D_\psi(x||\mu) = 0$ if and only if $x = \mu$, maximum likelihood for mean-parametrized vMF distributions (and exponential families in general) coincides with simple moment matching

$$\mu_{ML} = \operatorname*{argmax}_\mu \mathrm{LL}(\mu) = \frac{1}{N}\sum_{n=1}^{N} x^n.$$

To evaluate the learned density $p(x|\mu_{ML})$, however, we need either to evaluate $-D_\Psi(x||\mu) + \Psi(x)$ and hence both $\Psi(\mu)$ and $\nabla\Psi(\mu)$, or alternatively to transform mean parameters μ_{ML} into natural parameters $\eta_{ML} = \nabla\Psi(\mu_{ML})$ and evaluate $p(x|\eta_{ML})$.

Unfortunately, $\Psi = \Phi^*$ and its gradient $\nabla\Psi(\mu)$ are not available in closed form. Since $\nabla\Psi = (\nabla\Phi)^{-1}$ by the properties of the Legendre transform (Bauschke & Borwein, 1997), practical maximum likelihood for vMF distributions often depends on numerical root finders for $||\nabla\Phi(\eta) - \mu_{ML}||$ to obtain maximum-likelihood estimates η_{ML}.

In the following we derive a second-order ordinary differential equation whose solution under appropriate boundary conditions gives both $\Psi(\mu)$ and $\nabla\Psi(\mu)$ at a precision limited only by the numerical ODE solver. We furthermore give simple approximations to $\nabla\Psi(\mu)$ and $\Psi(\mu)$ based on rational functions and their antiderivatives. We provide code in Python [2].

2. Related work

The duality of natural and mean parametrizations for exponential families was explored in depth by Banerjee et al. (2005b), who presented a general expectation maximization (Dempster et al., 1977) algorithm for mixture models with exponential family mixture components in mean parametrization. This EM variant, dubbed 'Bregman clustering' due to reliance on $D_\Psi(x||\mu) - \Psi(x)$, was however not applicable to vMF mixture components because the vMF negative entropy Ψ and associated Bregman divergence were not available in closed form.

For the special case of vMF mixture components, Banerjee et al. (2005a) in the same year instead presented an adjusted EM algorithm operating in natural parameter space, and applied it to document clustering problems in high-dimensional feature spaces. This EM algorithm for mixtures of vMF ('MovMF') distributions included an approximation to the mapping from (weighted) average data norms $||\frac{1}{N}\sum_n x^n|| \approx ||\mu||$ to vMF concentration parameters $\kappa = ||\eta||$ as needed to find the maximum likelihood natural parameters, thus approximating a scalar mapping $\psi'(||\mu||) = ||\eta||$. They discussed further refinements to their approximation, but did not discuss consequences for estimating $\Psi(\mu)$ or $D_\Psi(x||\mu)$, which would have made it possible to also use Bregman clustering in the MovMF case. Since then, additional studies have developed methods to numerically evaluate the mapping $||\mu|| \mapsto ||\eta||$ using iterative solvers (Tanabe et al., 2007; Sra, 2012; Song et al., 2012; Hornik & Grün, 2014; Christie, 2015; Hasnat et al., 2016).

A separate line of work in theoretical statistics investigates variance functions $V(\mu)$ to characterize natural exponential families (Morris, 1983; Letac & Mora, 1990; Bar-Lev et al., 1994; Ghribi et al., 2015). Variance functions describe the (co-)variance $\mathrm{Cov}[X]$ of a random variable X in terms of its mean $\mathbb{E}[X] = \mu$, and thus are related to negative entropies via the inverse Hessian, $V(\mu) = \left(\nabla^2\Psi(\mu)\right)^{-1}$. To the best of our understanding, the variance function of the vMF distribution has not thus far been characterized. As we will show, it can be computed numerically by solving a second-order univariate ODE.

3. Results

We derive a second-order ODE to describe $\psi''(||\mu||)$, the second derivative of the scalar-valued radial profile $\psi(||\mu||)$ of Ψ. Note that the negative entropy inherits radial symmetry from $\Phi(\eta) = \phi(||\eta||)$:

$$\begin{aligned}\Psi(\mu) &= \max_\eta \mu^\top\eta - \Phi(\eta) = \max_\eta \mu^\top\eta - \phi(||\eta||)\\ &= \max_{||\eta||} ||\mu||\cdot||\eta|| - \phi(||\eta||)\\ &= \psi(||\mu||).\end{aligned}$$

The radial profile $\psi : [0, 1[\to \mathbb{R}_{\geq 0}$ is sufficient to also describe the gradients $\nabla\Psi(\mu)$, since their direction is always aligned with μ. Similarly, the Hessian $\nabla^2\Psi(\mu) \in \mathbb{R}^{D\times D}$ can be expressed in terms of $\psi'(||\mu||)$ and $\psi''(||\mu||)$ (see appendix A.1).

3.1. Covariances and support on the unit hypersphere

For natural exponential families with associated log-partition function Φ and negative entropy Ψ, the inverse function theorem applied to pairs (μ, η) with $\mu = \nabla\Phi(\eta)$ gives

$$\nabla^2\Psi(\mu) = \left(\nabla^2\Phi(\eta)\right)^{-1} = \left(\mathrm{Cov}[X|\mu]\right)^{-1}.$$

We can derive an important property of the vMF variance function $V(\mu) = \mathrm{Cov}[X|\mu]$ simply from the fact that the support is restricted to S^{D-1}: Any random variable X with

[2] See https://github.com/vmf-negentropy/vmf-negentropy

support limited to the unit ball $||X|| \leq 1$ has an upper bound on its variances and hence on the eigenvalues of $\mathrm{Cov}[X]$ in terms of the mean. For random variables with $\mathbb{E}[X] = \mu$ and support limited to the unit ball, we have (see appendix A.2)

$$tr\left(\mathrm{Cov}[X|\mu]\right) \leq 1 - \mu^\top \mu.$$

Furthermore, equality is obtained if and only if all the probability mass lies on the surface of unit ball (i.e. on the hypersphere),

$$tr\left(\mathrm{Cov}[X|\mu]\right) = 1 - \mu^\top \mu \;\; \text{iff} \;\; ||X|| = 1 \;\; \text{a.s.}$$

Note that this latter result is based only on the support. It does not depend on the distribution of mass over S^{D-1}.

3.2. Radial symmetry and a second-order ODE for the von Mises-Fisher negative entropy

From the eigendecomposition $\nabla^2\Psi(\mu) = U\Lambda U^\top$, we have

$$tr\left(\nabla^2\Psi(\mu)^{-1}\right) = \sum_i \Lambda_{ii}^{-1} = \sum_i \frac{1}{\lambda_i}$$

for eigenvalues $\lambda_i(\mu) > 0$ of $\nabla^2\Psi(\mu)$. For radially symmetric $\Psi(\mu) = \psi(||\mu||)$, we have (see appendix A.1)

$$|\nabla^2\Psi(\mu)| = \left(\frac{\psi'(||\mu||)}{||\mu||}\right)^{D-1} \psi''(||\mu||),$$

where $|\cdot|$ on the left-hand side denotes the determinant. The eigenvalues of the Hessian are $\lambda_1 = \psi''(||\mu||)$ (with eigenvector μ) and $\lambda_i = \frac{\psi'(||\mu||)}{||\mu||}, i = 2, \ldots, D$.

Thus for the von Mises-Fisher distribution $p(x|\mu)$ with support over S^{D-1} and radially symmetric $\Psi(\mu)$, we find

$$\begin{aligned} tr\left(\mathrm{Cov}[X|\mu]\right) &= \sum_i \lambda_i^{-1} = \frac{1}{\psi''(||\mu||)} + (D-1)\frac{||\mu||}{\psi'(||\mu||)} \\ &= 1 - ||\mu||^2. \end{aligned}$$

Reordering terms thus yields a second-order nonlinear ODE

$$\psi''(||\mu||) = \frac{\psi'(||\mu||)}{(1 - ||\mu||^2)\psi'(||\mu||) + (1-D)||\mu||}.$$

This result is exact, in that no approximations were used to arrive at this identity. We use the standard Runge-Kutta 4(5) numerical solver for initial value problems from the SciPy package (Virtanen et al., 2020) to numerically integrate this second-order ODE from specified initial conditions, with absolute and relative error tolerances of 10^{-12}.

3.3. Initial conditions

Since the base measure is radially symmetric on the hypersphere, the vMF distribution with $\eta = 0$ has $\mu = 0$, and by Legendre duality, $\nabla\Psi(0) = 0$. Thus we have

$$\psi(0) = \lim_{\eta \to 0} -\Phi(||\eta||) = (D/2 - 1)\log(2) + \log\Gamma(D/2).$$

However, setting $\psi'(0) = ||\nabla\Psi(0)|| = 0$ in the ODE for $||\mu|| = 0$ yields an ill-defined initial condition. In practice we approximate $\psi'(0) = \lim_{\epsilon \downarrow 0} \epsilon$ with a small value ($\leq 10^{-4}$, see appendix A.3 for an evaluation of this). Knowing the value $\psi'(||\mu||)$ for any $||\mu|| > 0$ with sufficient precision would allow us to circumvent this issue by starting the integration from that point.

3.4. Connection to variance functions

Variance functions $V(\mu)$ have been recognized as a useful tool to (uniquely) characterize natural exponential families (Morris, 1983). The vMF variance function for given radial profiles ψ', ψ'' is (see appendix A.1)

$$V(\mu) = \frac{||\mu||}{\psi'(||\mu||)} I_D + \left(\frac{1}{\psi''(||\mu||)} - \frac{||\mu||}{\psi'(||\mu||)}\right)\frac{\mu\mu^\top}{||\mu||^2},$$

the trace of which recovers the result in section 3.2. Traces of variance functions have been studied as potential identifiers of natural exponential families (Arab et al., 2015). Due to the occurrence of ψ', ψ'', this description of the variance function is not closed-form.

We note that the general form of this variance function, a scaled identity matrix plus a scaled rank-one matrix $\mu\mu^\top$, also holds for other radially symmetric negative entropies with domain $||\mu|| \leq 1$, but whose associated exponential families unlike the vMF have support within the interior of the unit ball $||X|| \leq 1$, and thus for which $tr(V(\mu)) < 1 - \mu^\top\mu$ for at least some μ.

We further note that computing the negative entropy $\Psi(\mu)$ from (the reciprocal of) the variance function $(V(\mu))^{-1} = \nabla^2\Psi(\mu)$ involves finding a second anti-derivative. Even if the variance function were known in closed form, analytic anti-derivatives may not be available. Thus a numerical approach might remain necessary even in this case, e.g. through iterated line integrals starting from a point μ_0 with known $\Psi(\mu_0), \nabla\Psi(\mu_0)$ – for $\mu_0 = 0$, this becomes very similar to the numerical solution of our second-order ODE. However, the scalar form of our ODE comes from being based on a matrix trace, rather than from line integrals.

3.5. Closed-form approximation

Banerjee et al. (2005a) gave an excellent approximation to the derivative of the negative entropy profile

$$\psi'_B(||\mu||) := \frac{||\mu||(D - ||\mu||^2)}{1 - ||\mu||^2} \approx \psi'(||\mu||)$$

based on a first-order truncation of a continued fraction series representation of $\phi'(||\eta||)$ (Watson, 1995), with an

additive error correction term derived from post-hoc inspection. The result is useful for a wide range of dimensionalities D. Their main use was to approximate the mean-to-natural parameter mapping $\eta = \nabla\Psi(\mu) = \frac{\psi'(||\mu||)}{||\mu||}\mu$ that occurs in maximum likelihood and expectation maximization for exponential families in natural parametrization. A question when wanting to work with the mean parametrization throughout is if the anti-derivative

$$\psi_B(||\mu||) = \frac{1}{2}||\mu||^2 + \frac{1-D}{2}\log\left(1-||\mu||^2\right)$$

can serve as a basis to approximate $\psi(||\mu||) = \Psi(\mu)$.

As we will show, this anti-derivative is indeed a plausible approximation to the numerical solutions for ψ obtained from our ODE approach. Based on this finding, and with the second-order ODE at hand, we propose an augmented numerical approximation obtained by refining ψ_B:

$$\begin{aligned}\tilde{\psi}_1'(||\mu||) &= \frac{(D-1)||\mu||}{1-||\mu||^2 - 1/\psi_B''(||\mu||)} \\ &= \frac{(1-D)\left(||\mu||^5 + (D-3)||\mu||^3 + D||\mu||\right)}{||\mu||^6 + (D-3)||\mu||^4 + ||\mu||^2 + D - 1},\end{aligned}$$

which has a closed-form anti-derivative

$$\begin{aligned}\tilde{\psi}_1(||\mu||) = &\frac{(1-D)}{2}\log(1-||\mu||^2) \\ &+ \frac{1-D}{4s}\left(\log(v+||\mu||^2+s) - \log(v+||\mu||^2-s)\right) \\ &+ const.\end{aligned}$$

for $v = \frac{D}{2} - 1$ and $s = \sqrt{\left(\frac{D}{2}-1\right)^2 - D}$. A further refinement can be achieved with

$$\tilde{\psi}_2'(||\mu||) = \frac{(D-1)||\mu||}{1-||\mu||^2 - 1/\tilde{\psi}_1''(||\mu||)},$$

yielding a rational function of degree ten (see appendix A.4). Computing the antiderivative of $\tilde{\psi}_2'$ involves solving a fifth-order polynomial with coefficients given by D, making it somewhat cumbersome. We hence use $\tilde{\psi}_1$ in the following.

Since $\tilde{\psi}_r'', r \in \{1, 2\}$ are also available in closed form, we can reformulate the ODE to numerically solve for $\psi - \tilde{\psi}_r$ and $\psi' - \tilde{\psi}_r'$ (see A.4), or for large $D \gg 10$ omit numerical integration of the ODE entirely and simply approximate $\psi \approx \tilde{\psi}_1, \psi' \approx \tilde{\psi}_2'$.

4. Numerical experiments

4.1. Direct numerical verification

To assess the accuracy of $\psi(||\mu||) = \Psi(\mu)$ and $\psi'(||\mu||) = ||\nabla\Psi(\mu)||$ obtained by numerical integration of the second-order ODE we adopted the following procedure.

We began by choosing a set of natural parameters η of different magnitudes. These were used to evaluate negative entropies

$$-H[X|\eta] = ||\eta||\phi'(||\eta||) - \phi(||\eta||) - \frac{D}{2}\log 2\pi,$$

and the corresponding mean parameters μ by numerical calculation of the roots of

$$\|\nabla\Phi(\eta)\| - ||\mu|| = \phi'(||\eta||) - ||\mu|| = \frac{I_{D/2}(||\eta||)}{I_{D/2-1}(||\eta||)} - ||\mu||$$

with the Newton-Raphson method (see appendix A.5 for details). This procedure yielded triples $(\mu, \nabla\Psi(\mu) = \eta, \Psi(\mu) = -H[X|\eta])$ based on prior knowledge of η. We then asked whether we could recover Ψ and its gradient starting from μ alone.

In fact, this process is complicated by the need to evaluate the modified Bessel function for orders $D/2$, and $D/2 - 1$ when both D is large. Fast and asymptotically valid approximations for large $||\eta||$ have been known for long (Abramowitz & Stegun, 1968), but in many applications with high dimensional data, we need to evaluate both narrow and broad vMF distributions. Several previous studies on parameter inference in vMF distributions found it necessary to develop their own approximations of $\log I_v$ and I_v/I_{v-1} needed to compute the log-partition function $\Phi(\eta)$ and its gradient (Tanabe et al., 2007; Sra, 2012; Song et al., 2012; Hornik & Grün, 2014). Here, we truncated a continuous fraction representation of I_v/I_{v-1} (see appendix A.5) to obtain an approximation of the ratio of modified Bessel functions. For $\log I_v$, we followed recent work on vFM modeling (Kim, 2021) and used a library for arbitrary-precision floating point arithmetic (Johansson et al., 2013), which we found sufficient for the dimensionalities D explored here.

When evaluating the negative entropy and its gradient at several points μ_k, we sorted them by norm and solve the ODE from $||\mu|| = 0$ to $||\mu|| = \text{argmax}_k\, ||\mu_k||$, forcing the solver to stop at all intermediate $||\mu_k||$. This way we only needed to solve the ODE once. To speed up subsequent evaluations after a first solve, one can also store checkpoints of the ODE solutions at intermediate $||\mu||$ for use as initial conditions.

As seen in Fig. 1, the ODE solutions agree well with the η-derived values of both the negative entropy and its gradient.

We next compared the various closed-form approximations to ψ, ψ' (Section 3.5) and the resulting estimate of $D_\Psi(0||\mu) = \nabla\Psi(\mu)^\top\mu - \Psi(\mu)$ against numerical solutions of the ODE (Fig. 2). We computed both the absolute error magnitude $|f(\mu) - \hat{f}(\mu)|$ and the relative error $|f(\mu) - \hat{f}(\mu)|/f(\mu)$ for different values of μ and D, where

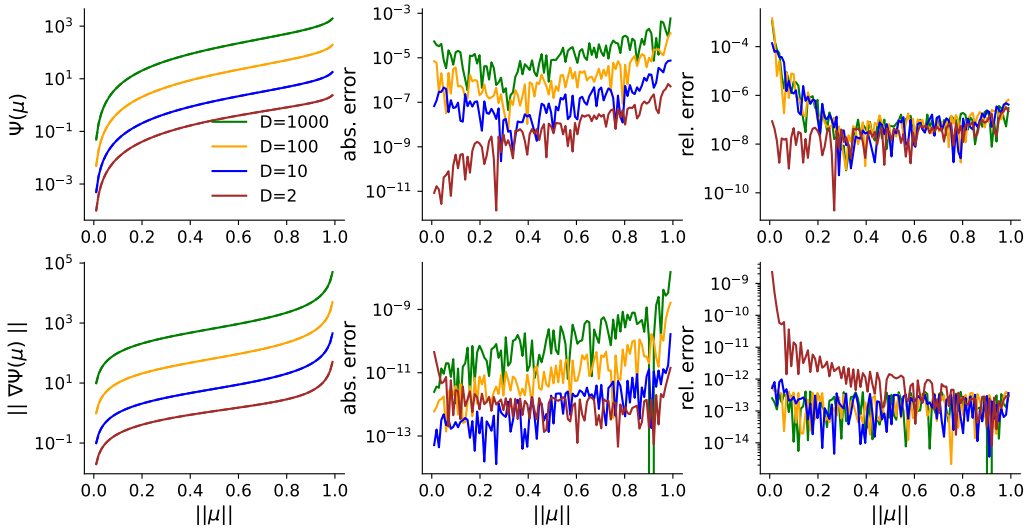

Figure 1. Numerical comparison of ODE results for the negative entropy of the vMF. Top: comparison of negative entropy computed from second-order ODE and von Mises Fisher entropy $H[p|\eta]$ computed in natural parametrization. Note that the shown errors include those from mapping between natural- and mean-parameter spaces. Bottom: gradient of the negative entropy. Since the direction of the gradient is given by μ, we only need to identify $||\nabla\Psi(\mu)|| = \psi'(\mu)$. We compare the numerical solution of the second-order ODE against root-finding results (Newton-Raphson on $\phi'(||\eta||) - ||\mu||$).

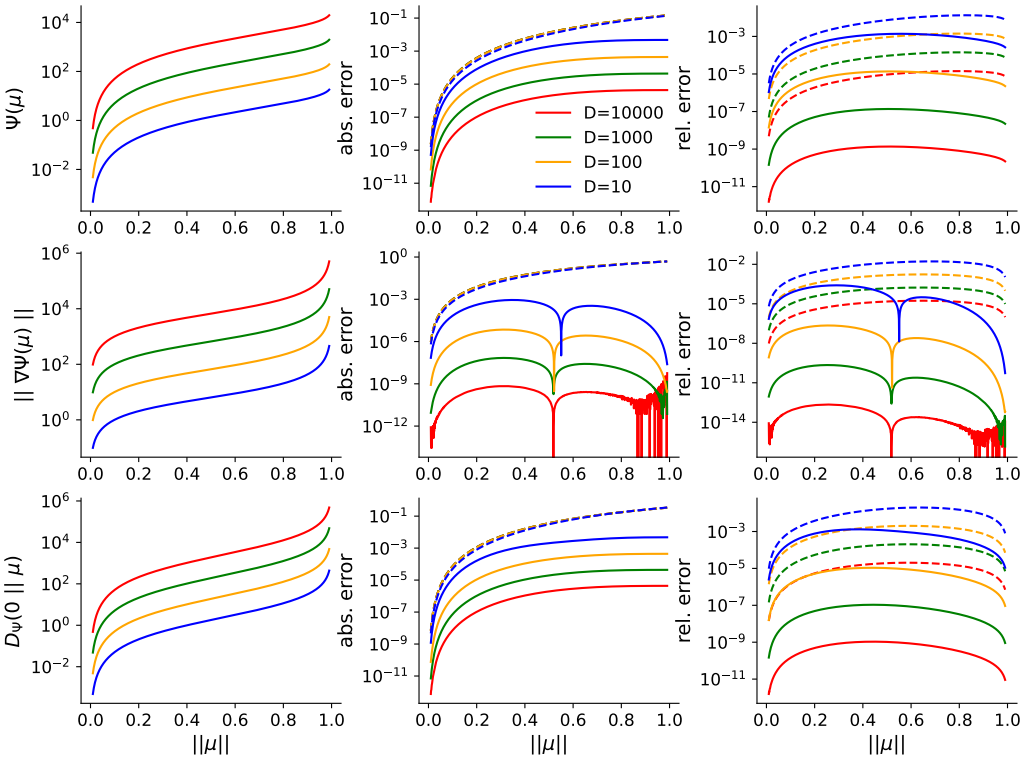

Figure 2. Evaluation of closed-form approximations for $\Psi(\mu) = \psi(||\mu||)$, $||\nabla\Psi(\mu)|| = \psi'(||\mu||)$, and $D_\Psi(0||\mu) = ||\mu||\psi'(||\mu||) - \psi(||\mu||)$, against numerical ODE solution, for different dimensions D. Left: numerical solutions to the second-order ODE in dimension $D = 10^i$, $i \in \{1, 2, 3, 4\}$. Graphs for the numerical approximations are visually indistinguishable. Center: absolute errors between ODE solution and closed-form approximations. Right: relative errors between ODE solution and closed-form approximations. Dashed lines for approximations $\hat{\Psi}_B$ based on Banerjee et al. (2005), solid lines for our adjusted approximations $\tilde{\Psi}$ (see text). Dents in relative errors for the gradient norms come from the closed-form approximations switching between under- and overestimation.

$f \in \{\psi, \psi', D_\Psi(0||\cdot)\}$ and $\hat{f}$ is the corresponding estimate based on ψ_B or $\tilde{\psi}_r$.

While the approximation ψ_B given by Banerjee et al. (dashed lines) is quite accurate, both the absolute and (to a lesser degree) relative errors tend to increase with $||\mu||$. By contrast, the absolute error appears to depend only weakly on dimension $D \in \{10, 100, 1000, 10000\}$, even though the values of all three of the target functions increase by several orders of magnitude from $D = 10$ to $D = 10^4$. As a result, the relative errors fell with D. Our adjusted closed-form estimate $\tilde{\psi}_1$ and our estimate $\tilde{\psi}_2'$ for the derivative (solid lines) achieve lower absolute and relative errors than ψ_B accrosss the full range of investigated $||\mu||$ (Fig. 2). Interestingly, also the absolute errors for our approximations shrink with increasing dimensionality D. For $D \geq 50000$, we noticed little difference in absolute error to the ODE solution between the approximation $\tilde{\psi}_2'(||\mu||) = \kappa$ and root-finding on $\phi'(\kappa) - ||\mu||$ via Newton-Raphson and common numerical approximations to $\phi' = \frac{I_{D/2}}{I_{D/2-1}}$ (see appendix A.5).

4.2. Bregman clustering

Bregman clustering (Banerjee et al., 2005b) is based on maximum-likelihood fitting of a mixture exponential family model given by

$$p(x \mid \theta = \{\pi_k, \mu_k\}_{k=1}^K) = \sum_{k=1}^K \pi_k \, p(x \mid \mu_k),$$

with $\pi_k \geq 0, \sum_k \pi_k = 1$ and exponential family $p(x|\mu)$ defined in mean parametrization. Fitting proceeds by expectation maximization, using $\log p(x|\mu) = -D_\Psi(x|\mu) + f(x)$ for efficient likelihood evaluation based on the negative entropy Ψ. Here, we exploited our novel formulation of the vMF negative entropy and derivatives to apply Bregman clustering to the problems of clustering documents (Banerjee et al., 2005a) from the 'news20' and 'classic3' datasets (Dhillon et al., 2003; Lang, 1995), which each comprise several thousand documents from multiple categories. Previous models of these data have relied on natural parameter vMF formulations, as methods to evaluate Ψ were unavailable.

Documents indexed $n = 1, \ldots, N$ were mapped onto feature representations $x^n \in S^{D-1}$ using the term-frequency inverse document frequency (Robertson, 2004) representation. D is given by the size of the dataset-derived vocabulary that defines the word-frequency embedding space. The vocabularies are large, and so the resulting representation vectors are high-dimensional with D ranging in the thousands to tens of thousands. Following Banerjee et al. (2005a), we also created smaller datasets by subsampling 100 documents per category. See appendix A.6 for further information on datasets and feature spaces, and Table 1 for a summary of dataset features.

Table 1. Datasets for document clustering. Number of categories K_{true}, number of documents N and feature space dimension D.

DATA SET	K_{true}	N	D
CLASSIC3	3	3891	4674
CLASS300	3	300	2551
NEWS20	20	18803	28571
NEWS20-SMALL	20	2000	16218

We fitted MovMF models with different numbers of mixture components and algorithmic variants to each dataset. Algorithmic variants were soft- and hard-assignment EM. Soft assignment EM computes the full cluster "responsibilities" $p(k|x, \theta)$ in the E-step, whereas in the hard-assignment E-step we assign each datapoint x^n to a single mixture component $k'_n = \text{argmax}_k \, p(k|x^n, \theta)$. For both soft- and hard assignment variants, we fitted a MovFM model in natural parametrization $\theta = (\pi, \{\eta_k\}_k)$ and another one in mean parametrization $\theta = (\pi, \{\mu_k\}_k)$, from the equivalent initialization $\mu_k^{\text{init}} = \nabla\Phi(\eta_k^{\text{init}})$ using 100 EM iterations. We compared results across 10 different random initializations. Bregman clustering fits used the closed-form approximation to ψ, ψ' described in section 3.5. Since it was previously noted that MovMF models applied to document clustering perform better with tied concentration parameters $||\eta_k|| = ||\eta_l||$, resp. $||\mu_k|| = ||\mu_l||$, for all $k, l = 1, \ldots, K$ (Zhong & Ghosh, 2003) , we also included such restricted model variants (Fig. 3, dashed lines).

We evaluated results using normalized mutual information between the true document topics and the clustering obtained from the responsibilities $\text{argmax}_k \, p(k|x^n, \theta_{ML})$. We did not find qualitative differences in the results between the corresponding mean- and natural-parameter models and algorithms, showcasing that Bregman clustering on high-dimensional hyperspheres is feasible with our approximations to the negative entropy and its gradient.

5. Discussion

We derived a second-order differential equation to describe the negative entropy of the von Mises-Fisher distribution and its associated variance function. To arrive at this result, we combined two features of the vMF distribution: that the base measure and consequently the negative entropy are radially symmetric, and that the base measure only has support on the boundary of the mean-parameter domain, i.e. the unit hypersphere.

Many other radially symmetric negative entropy functions with domains $||\mu|| \leq 1$ may exist and describe valid natural exponential families over the unit ball. These distributions would be described by different radial profiles $p(||x||)$, but

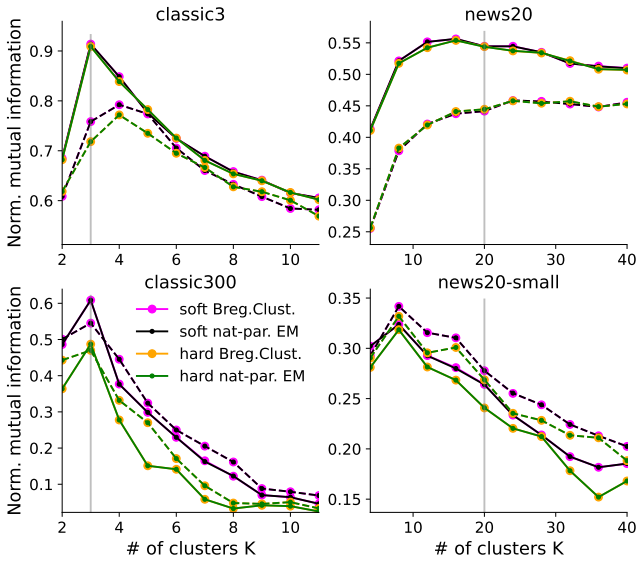

Figure 3. Clustering on the unit hypersphere. Normalized mutual information scores between true and estimated clusterings for two different variants of the document collections 'classic3' and '20newsgroup'. Documents are mapped by term-frequency inverse document frequency onto unit hyperspheres in dimensions $D = 4674$ for 'classic3' and $D = 28571$ for 'news20'. Gray lines mark true number of clusters. 'classic300' and 'news20-small' are subsampled datasets (see text). We compare several EM algorithms using mixture of vMF models. Results are averaged over 10 random seeds. Solid lines are for tied cluster concentration parameters, i.e. $||\eta_k|| = ||\eta_l||$ resp. $||\mu_k|| = ||\mu_l||$ for all $k, l = 1, \ldots, K$. Dashed lines are for separate concentration parameters. We find no qualitative difference between the results of the EM algorithms working in natural parameter space (black & green), and the Bregman clustering algorithms (magenta & orange) that work in mean parametrization using our numerical approximations to the Bregman divergence.

share the general form for the variance function of a scaled identity matrix plus rank one matrix. Conversely, other negative entropies with $tr(V(\mu)) = 1 - ||\mu||^2$ could be found which are not radially symmetric, but still describe natural exponential families with support over the unit hypersphere. Conditions on matrix-valued functions to be valid second derivatives of strictly convex functions have been investigated in machine learning (Richter-Powell et al., 2021) and statistics (Ghribi et al., 2015), and could be combined with the trace condition to describe new natural exponential families on the unit hypersphere.

Another notable aspect of negative entropies with $tr\left(\nabla^2\Psi(\mu)^{-1}\right) = 1 - ||\mu||^2$ is that they also describe exponential families that are not natural, in the sense that there is a $(D-1)$-dimensional pullback measure on the hypersphere. In the case of $D = 2$, that is the well-known univariate von Mises distribution over $[0, 2\pi[$. This is achieved here through forcing the base measure support onto the boundary of the mean-parameter domain as we find the maximal variance compatible with the domain shape. This is of potential interest beyond the vMF distribution because mean-parameter domains are interpretable and often known. In principle, larger gaps between the number of mean parameters D and the number of variables in the pullback measure could be achieved if the base measure support were further restricted into even lower-dimensional subsets of the domain boundary, such as a one-dimensional set of extreme points on a $(D-1)$-dimensional boundary.

We furthermore described fast approximations to the negative entropy and its first two derivatives that over a wide range of dimensionalities become better with higher dimensions. Importantly, these closed-form approximations are numerically cheap compared to iterative schemes based on numerical approximations to the vMF log-partition. In particular for small dimensions it however remains an open question how best to use the derived relations among derivatives of the negative entropy, without having to numerically solve the ODE at high precision.

References

Abramowitz, M. and Stegun, I. A. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1968.

Arab, T. B., Masmoudi, A., and Zribi, M. Trace of the variance-covariance matrix in natural exponential families. *Communications in Statistics-Theory and Methods*, 44(6):1241–1254, 2015.

Banerjee, A., Dhillon, I. S., Ghosh, J., Sra, S., and Ridgeway, G. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005a.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. Clustering with Bregman divergences. *Journal of machine learning research*, 6(10), 2005b.

Bar-Lev, S. K., Bshouty, D., Enis, P., Letac, G., Lu, I.-L., and Richards, D. The diagonal multivariate natural exponential families and their classification. *Journal of Theoretical Probability*, 7:883–929, 1994.

Bauschke, H. H. and Borwein, J. M. Legendre functions and the method of random Bregman projections. *Journal of convex analysis*, 4(1):27–67, 1997.

Bisson, G. and Hussain, F. Chi-sim: A new similarity measure for the co-clustering task. In *2008 Seventh International Conference on Machine Learning and Applications*, pp. 211–217. IEEE, 2008.

Christie, D. Efficient von Mises-Fisher concentration parameter estimation using Taylor series. *Journal of Statistical Computation and Simulation*, 85(16):3259–3265, 2015.

Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T., and Tomczak, J. M. Hyperspherical variational auto-encoders. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pp. 856–865. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Dhillon, I. S., Mallela, S., and Modha, D. S. Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 89–98, 2003.

Gautschi, W. and Slavik, J. On the computation of modified bessel function ratios. *Mathematics of Computation*, 32 (143):865–875, 1978.

Ghribi, A., Kokonendji, C. C., and Masmoudi, A. Characteristic property of a class of multivariate variance functions. *Lithuanian Mathematical Journal*, 55:506–517, 2015.

Hasnat, M. A., Alata, O., and Tremeau, A. Model-based hierarchical clustering with Bregman divergences and Fishers mixture model: application to depth image analysis. *Statistics and Computing*, 26:861–880, 2016.

Hornik, K. and Grün, B. On maximum likelihood estimation of the concentration parameter of von Mises-Fisher distributions. *Computational statistics*, 29:945–957, 2014.

Johansson, F., Steinberg, V., Kirpichev, S. B., Kuhlman, K. L., Meurer, A., Čertík, O., Van Horsen, C., Masson, P., Arias de Reyna, J., Hartmann, T., et al. mpmath: a python library for arbitrary-precision floating-point arithmetic. *Zenodo*, 2013.

Kim, M. On pytorch implementation of density estimators for von Mises-Fisher and its mixture. *arXiv preprint arXiv:2102.05340*, 2021.

Lang, K. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pp. 331–339. Elsevier, 1995.

Letac, G. and Mora, M. Natural real exponential families with cubic variance functions. *The Annals of Statistics*, 18(1):1–37, 1990.

Lewis, D. D., Yang, Y., Russell-Rose, T., and Li, F. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

Mardia, K. V. Statistics of directional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 37(3):349–371, 1975.

Mitchell, T. M. *Machine learning*. McGraw-Hill, 1997.

Morris, C. N. Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*, pp. 515–529, 1983.

Rao, C. R. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

Richter-Powell, J., Lorraine, J., and Amos, B. Input convex gradient networks. *arXiv preprint arXiv:2111.12187*, 2021.

Robertson, S. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

Rockafellar, R. T. *Convex analysis*, volume 11. Princeton University Press, 1997.

Salton, G. *The SMART retrieval system—experiments in automatic document processing*. Prentice-Hall, Inc., 1971.

Song, H., Liu, J., and Wang, G. High-order parameter approximation for von Mises-Fisher distributions. *Applied Mathematics and Computation*, 218(24):11880–11890, 2012.

Sra, S. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $i_s(x)$. *Computational Statistics*, 27:177–190, 2012.

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

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Wang, D., Cui, P., and Zhu, W. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1225–1234, 2016.

Watson, G. N. *A treatise on the theory of Bessel functions*. Cambridge University Press, 2nd Edition, 1995.

Xu, J. and Durrett, G. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4503–4513, 2018.

Zhong, S. and Ghosh, J. A comparative study of generative models for document clustering. In *Proceedings of the workshop on clustering high dimensional data and its applications in SIAM data mining conference*, 2003.

A. Appendix

A.1. Gradient and Hessian of radially symmetric functions

Let $f(x) = g(||x||)$ for $x \in \mathbb{R}^D$, $r = ||x|| \geq 0$ and I_D denote the $D \times D$ identity matrix. Then from basic calculus and the Sherman–Morrison formula,

$$
\begin{aligned}
(\nabla f)(x) &= \frac{g'(r)}{r} x, \\
(\nabla^2 f)(x) &= \frac{g'(r)}{r} I_D + \left(\frac{g''(r)}{r^2} - \frac{g'(r)}{r^3} \right) x x^\top, \\
(\nabla^2 f)^{-1}(x) &= \frac{r}{g'(r)} I_D + \left(\frac{1}{g''(r)} - \frac{r}{g'(r)} \right) \frac{x x^\top}{r^2}.
\end{aligned}
$$

For radially symmetric function $f(x)$ we furthermore have

$$
\begin{aligned}
|\nabla^2 f| &= \left| \nabla^2 f \right| \\
&= \left| \frac{g'(r)}{r} I_D + \left(\frac{g''(r)}{r^2} - \frac{g'(r)}{r^3} \right) x x^\top \right| \\
&\overset{*}{=} \left| \frac{g'(r)}{r} I_D \right| \left(1 + \frac{r}{g'(r)} \left(\frac{g''(r)}{r^2} - \frac{g'(r)}{r^3} \right) x^\top I_D x \right) \\
&= \left(\frac{g'(r)}{r} \right)^D \left(\frac{g''(r)}{r^2} \frac{r^3}{g'(r)} \right) \\
&= \left(\frac{g'(r)}{r} \right)^{D-1} g''(r),
\end{aligned}
$$

where the equality labelled by * derives from the matrix determinant lemma.

The eigenvalues indeed are $g''(r)$ for eigenvector x,

$$
\left(\nabla^2 f(x) \right) \; x = \frac{g'(r)}{r} x + \left(\frac{g''(r)}{r^2} - \frac{g'(r)}{r^3} \right) r^2 x = g''(r) x,
$$

and $g'(r)/r$ with multiplicity $D-1$ for eigenvectors y from the orthogonal space with $x^\top y = 0$,

$$
\left(\nabla^2 f(x) \right) \; y = \frac{g'(r)}{r} y.
$$

A.2. Bound on covariance trace

Let X be a $D-$dimensional random vector X with mean $\mathbb{E}[X] = \mu$ and restricted support such that $||X|| \leq 1$ a.s. Then

$$
\begin{aligned}
tr\left(\mathrm{Cov}[X]\right) &= \sum_{d=1}^{D} \mathbb{E}[X_d^2] - \mathbb{E}[X_d]^2 \\
&= \mathbb{E}[\, ||X||^2 \,] - \mu^\top \mu \\
&\leq 1 - \mu^\top \mu,
\end{aligned}
$$

with equality iff $\mathbb{E}[\, ||X||^2] = 1$.

A.3. Dependence of ODE on exact initial conditions

Solving the ODE for the radial profiles $\psi(||\mu||), \psi'(||\mu||)$ from the center $||\mu|| = 0$ outwards is problematic due to the condition $\psi'(0) = \Psi(0) = 0$, which causes $\psi''(0)$ to become a fraction of zero over zero. We instead initialize $\psi'(0) > 0$ as a small value, and explore the dependence on $\psi'(0)$ in Fig. 4 for dimensions $D = 2$ and $D = 100$. In particular for larger D, the effects of different choices for $\psi'(0)$ on the full radial profile $\Psi(\mu) = \psi(||\mu||)$ are limited, because the first derivative converges quickly even from initializations that differ by four orders of magnitude. This attracting behavior of the solution for $||\nabla\Psi(\mu)|| = \psi'(||\mu||)$ in the forward direction also makes it very difficult to solve the ODE in reverse, e.g. from $||\mu|| = 0.5$ towards $||\mu|| = 0$. For numerical results in this study, we initialized the ODE at values between $\psi'(0) = 10^{-6}$ ($D = 2$) and $\psi'(0) = 10^{-4}$ ($D = 10000$).

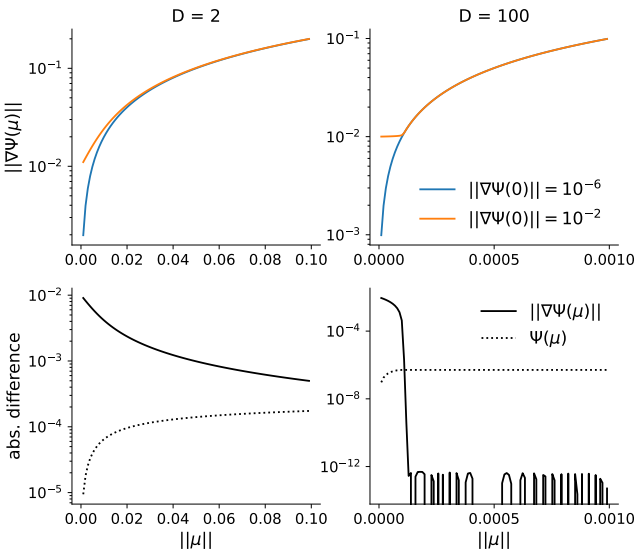

Figure 4. Dependence of ODE solutions on initial condition. Top: Numerical solutions $||\nabla\Psi(\mu)|| = \psi'(\mu)$ to second-order ODE solved from $\psi'(0) = 10^{-6}$ (blue) and from $\psi'(0) = 10^{-2}$ (orange). Solutions for first derivative ψ' approach one another quickly, in particular for larger dimension D. Bottom: resulting absolute difference between ODE solutions shown at the top, for radial profile of the negative entropy (dotted line), and the first derivative (solid line).

A.4. Closed-form approximation of negative entropy

We use the following closed-form approximation to the negative entropy of the vMF distribution in D dimensions,

along with its first two derivatives:

$$\tilde{\psi}_1(||\mu||) = \frac{(1-D)}{2}\log(1-||\mu||^2) + \frac{1-D}{4s}\left(\log(v+||\mu||^2+s) - \log(v+s)\right) - \frac{1-D}{4s}\left(\log(v+||\mu||^2-s) - \log(v-s)\right),$$

$$\tilde{\psi}_1'(||\mu||) = \frac{(D-1)||\mu||}{1-||\mu||^2} + \frac{(D-1)||\mu||}{||\mu||^4+(D-2)||\mu||^2+D-1},$$

$$\tilde{\psi}_1''(||\mu||) = (D-1)\frac{1+||\mu||^2}{(1-||\mu||^2)^2} + (D-1)\frac{-3||\mu||^4-(D-2)||\mu||^2+D-1}{(||\mu||^4+(D-2)||\mu||^2+D-1)^2},$$

where $v = \frac{D}{2} - 1$ and $s = \sqrt{v^2 - D}$. We include the second derivative here since it is needed to compute the vMF variance function.

We note that further refinements to the approximation to ψ' are possible through another iteration

$$\tilde{\psi}_2'(||\mu||) = \frac{(D-1)||\mu||}{1-||\mu||^2 - 1/\tilde{\psi}_1''(||\mu||)}, = \frac{(D-1)||\mu||}{1-||\mu||^2} + \frac{(D-1)||\mu||\alpha(||\mu||)^2}{(D(1+||\mu||^2)-2)\alpha(||\mu||)^2+\beta(||\mu||)},$$

where $\alpha(||\mu||) = (D-1)(1+||\mu||^2) - ||\mu||^2(1-||\mu||^2)$, $\beta(||\mu||) = (1-||\mu||^2)^2(-3||\mu||^4+(2-D)||\mu||^2+D-1)$. Beyond that, we find the next iteration $\tilde{\psi}_3'$ to do worse than $\tilde{\psi}_2'$ and fall between $\tilde{\psi}_1'$ and $\tilde{\psi}_2'$ in terms of approximation quality.

The integral representations of

$$\tilde{\psi}(||\mu||) = \int_0^{||\mu||} \tilde{\psi}'(r)dr,$$
$$\tilde{\psi}'(||\mu||) = \int_0^{||\mu||} \tilde{\psi}''(r)$$

are analogous to those of ψ and ψ', allowing to express

$$\psi(||\mu||) = \tilde{\psi}(||\mu||) + f(||\mu||),$$
$$\psi'(||\mu||) = \tilde{\psi}'(||\mu||) + f'(||\mu||),$$

where (f, f') is the solution to the second-order ODE arising from

$$f'' = \frac{\tilde{\psi}'(t) + f'}{(1-t^2)(\tilde{\psi}'(t)+f')) + (1-D)t} - \tilde{\psi}''(t).$$

For large D, we found the correction terms $f(||\mu||), f'(||\mu||)$ to become increasingly irrelevant, so one could try solving the ODE with lower precision, or omit it entirely ($f = f' = 0$).

A.5. Numerical comparison of gradient approximation.

We directly compare the numerical approximation $\tilde{\psi}_2'(\mu) \approx ||\nabla\Psi(\mu)|| = \kappa$ against root-fining on $\phi(\kappa) - \mu$, as is commonly used in maximum likelihood. We compare against truncated Newton-Raphson (Banerjee et al., 2005b; Sra, 2012), which requires a numerical evaluation of $\phi'(\kappa) = \frac{I_{D/2}(\kappa)}{I_{D/2-1}(\kappa)}$, which in turn is non-trivial for large D. We follow Hornik and Grün (2014) in truncating a Perron continuous fraction representation (Gautschi & Slavik, 1978) of

$$\frac{I_v(\kappa)}{I_{v-1}(\kappa)} = \frac{\kappa}{2v+\kappa - \frac{(2v+1)\kappa}{2v+1+2\kappa - \frac{(2v+3)\kappa}{2v+2+2\kappa-\ldots}}}.$$

This root finding scheme thus includes two nested loops: an outer loop over M Newton-Raphson steps, and an inner loop over L iterations of the continued fraction. The initialization is given by $\psi_B'(||\mu||)$ (Banerjee et al., 2005a). We found empirically that for $D > 10$ truncating the infinite continuous fraction after about $L = 20$ iterations gives good results up to numerical precision of double-precision floating point numbers. Every such iterations beyond the first two requires six algebraic operations. We could alternatively use a Gauss continuous fraction representation instead, which would incur five algebraic operations per iterations but be less stable (Hornik & Grün, 2014). Convergence to double precision typically requires $M = 2$ or $M = 3$ Newton-Raphson steps, for several hundred algebraic operations overall. Evaluating $\tilde{\psi}_2'(||\mu||)$ for given $D, ||\mu||$ in comparison requires 23 algebraic operations or less. We directly compare the two methods of estimating $\kappa = \psi'(||\mu||)$ for large D in figure 5.

A.6. Datasets and pre-processing

We consider the 'classic3' and 'news20' datasets previously used in the natural language processing literature (Mitchell, 1997; Bisson & Hussain, 2008; Wang et al., 2016) .
'classic3' comprises the $K_{true} = 3$ document collections CRANFIELD ($N = 1400$), CISI ($N = 1460$) and MEDLINE ($N = 1033$) of abstracts from scientific publications in engineering, information science and medical journals, respectively. The goal of clustering the documents is to assign the abstracts to their respective scientific field.
'20 newsgroups' ('news20') is a collection of 19997 messages from users on USENET newsgroups, distinguished by $K_{true} = 20$ different subject matters including religion, computer graphics and motorcycles. Removing duplicates, we are left with $N = 18803$ individual messages. We remove message headers to avoid the possibility of trivial assignment of messages to subject matter. The goal of clustering is to assign each document to its subject matter.
We also follow previous work (Banerjee et al., 2005a) in creating smaller versions of these datasets by sampling 100 documents per category from both 'classic3' and 'news20',

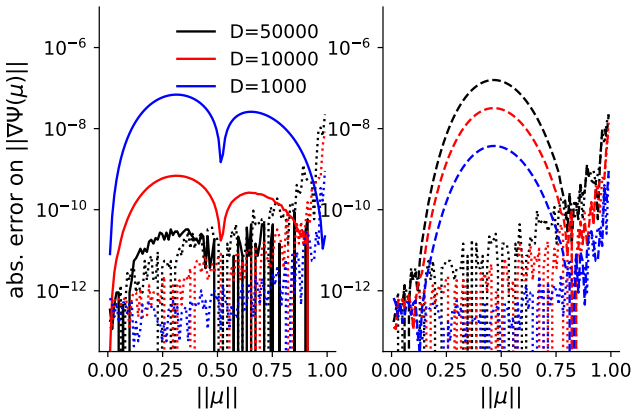

Figure 5. Direct comparison of the approximation $\tilde{\psi}_2'(||\mu||) \approx ||\nabla\Psi(\mu)|| = \kappa$ against truncated Newton-Raphson for finding the root of $\phi'(\kappa) - ||\mu||$. Left: Absolute errors $|\hat{\kappa} - ||\nabla\psi(\mu)||||$ with $||\nabla\psi(\mu)|| = \psi'(||\mu||)$ from the numerical solution of the second-order ODE and $\hat{\kappa}$ estimated via $\tilde{\psi}_2'$ (solid lines) and Newton-Raphson (dotted lines) with $M = 2$ steps and $L = 20$ iterations to estimate $\phi'(\kappa)$. Right: Same as left, but for $\hat{\kappa}$ being estimated with Newton-Raphson with either more (dotted, $M = 3$, $L = 50$) or less computations (dashed, $M = 2$, $L = 15$). At $D = 50.000$ (black), the tenth-degree rational function $\tilde{\psi}_2$ performs at around numerical precision (double-precision floating points), and thus comparable to the more expensive iterative root finder.

resulting in the much smaller 'classic300' and 'news20-small' datasets with perfectly balanced labels.

For each of these datasets, we create a feature mapping from documents onto feature representations on the unit hypersphere with term-frequency inverse term frequency, which starts out from the occurrence counts v_d^n of feature $d = 1, \ldots, D$ in document $n = 1, \ldots, N$. These term frequencies are weighted by the (log-) inverse document frequencies $g_d = \log(N/\sum_n v_d^n)$ and subsequently normalized per document to account for varying document lengths:

$$x^n = \frac{v^n \odot g}{||v^n \odot g||},$$

where $\odot$ refers to the Hadamard product.

The features indexed $d = 1, \ldots, D$ are tokens from a dataset-specific vocabulary. Vocabularies are built from all tokens occurring within a given dataset of N documents, which occur in at least P different documents (P=7/6/2/2 for 'news20'/'classic3'/'news20-small'/'classic300', respectively). We chose P to obtain vocabulary sizes D comparable to previous studies. We also remove overly common words appearing in more than 15% of documents, and remove any token that also appears on the SMART list (Salton, 1971) of common stopwords (see e.g. appendix 11 of Lewis et al. (2004)). Documents that do not contain a single token from the final vocabulary ($||v^n \odot g|| = 0$) are removed. This only happened for two documents from the CRANFIELD collection in the 'classic3' dataset.