Harmonic Exponential Families on Manifolds

Taco S. Cohen T.S.COHEN@UVA.NL
University of Amsterdam

Max Welling M.WELLING@UVA.NL
University of Amsterdam
University of California Irvine
Canadian Institute for Advanced Research

Abstract

In a range of fields including the geosciences, molecular biology, robotics and computer vision, one encounters problems that involve random variables on manifolds. Currently, there is a lack of flexible probabilistic models on manifolds that are fast and easy to train. We define an extremely flexible class of exponential family distributions on manifolds such as the torus, sphere, and rotation groups, and show that for these distributions the gradient of the log-likelihood can be computed efficiently using a non-commutative generalization of the Fast Fourier Transform (FFT). We discuss applications to Bayesian camera motion estimation (where harmonic exponential families serve as conjugate priors), and modelling of the spatial distribution of earthquakes on the surface of the earth. Our experimental results show that harmonic densities yield a significantly higher likelihood than the best competing method, while being orders of magnitude faster to train.

1. Introduction

Many problems in science and engineering involve random variables on manifolds. In the geosciences, for example, one deals with measurements such as the locations of earthquakes and weather events on the spherical surface of the earth. In robotics and computer vision, unobserved Lie group transformations such as rotations and rigid-body motions play an important role in motion understanding, localization and alignment problems. Classical probabilistic models cannot be applied to data on manifolds because these models do not respect the manifold topology (having discontinuities at the value $0 \equiv 2\pi$ of a circular variable, for instance), and because they are not equivariant (a manifold-preserving transformation of the data would take the distribution outside the model family). Among manifold distributions there are currently none that are both flexible and efficiently trainable.

In this paper we study a very flexible class of densities on compact Lie groups (such as the group of rotations in two or three dimensions) and homogeneous spaces (such as the circle, torus, and sphere). We refer to these densities as harmonic exponential families, because they are based on a generalized form of Fourier analysis known as non-commutative harmonic analysis (Chirikjian & Kyatkin, 2001). Specifically, the sufficient statistics of these families are given by functions that are analogous to the sinusoids on the circle.

We show that the moment map (Wainwright & Jordan, 2007) that takes the natural parameters of an exponential family and produces the moments of the distribution can be computed very efficiently for harmonic exponential families using generalized Fast Fourier Transform (FFT) algorithms. This leads directly to a very efficient maximum-likelihood estimation procedure that is applicable to any manifold for which an FFT algorithm has been developed, and that enjoys the convexity and convergence properties of exponential families.

We apply the harmonic exponential family of the sphere to the problem of modeling the spatial distribution of earthquakes on the surface of the earth. Our model significantly outperforms the best competing method (a mixture model), while being orders of magnitude faster to train.

Harmonic exponential families also arise naturally as conjugate priors in a Bayesian framework for estimating transformations from correspondence pairs. In this case the points on the manifold (the transformations) are not observed directly. Instead we see a pair of vectors x and y

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

— images, point clouds, or other data — that provide information about the transformation that produced one from the other. If we have a prior $p(g)$ over transformations and a likelihood $p(y \mid x,\, g)$ that measures how likely y is to be the g-transformed version of x, we can consider the posterior over transformations $p(g \mid x,\, y)$. Typically, the posterior turns into a complicated and intractable distribution, but we show that if the prior is a harmonic density and the likelihood is Gaussian, the posterior distribution is again a harmonic density whose parameters are easily obtained using the generalized FFT algorithm. Furthermore, the global mode of this posterior (the optimal transformation) can be computed efficiently by performing yet another FFT.

In this paper we provide both an abstract treatment of the theory of harmonic densities that covers a fairly broad class of manifolds in a uniform and easily understandable manner, and a concrete instantiation of this theory in the case of the rotation groups $\mathrm{SO}(2)$ and $\mathrm{SO}(3)$, and their homogeneous spaces, the circle S^1 and sphere S^2.

The rest of this paper is organized as follows. We begin by discussing related work in section 2, followed by a brief summary of non-commutative harmonic analysis for a machine learning audience in section 3. In section 4, we define harmonic exponential families and present an FFT-based maximum-likelihood estimation algorithm. In section 5, we show that harmonic exponential families are the conjugate priors in the Bayesian transformation estimation problem, and present an efficient MAP inference algorithm. Our earthquake modelling experiments are presented in section 6, followed by a discussion and conclusion.

2. Related Work

The harmonic exponential families were first defined abstractly (but not named) by Diaconis (1988). Despite their long history and a significant body of literature devoted to them (see Mardia & Jupp (1999)), this class of densities has remained intractable until now for all but the simplest cases.

In fact, various commonly used distributions on the circle and the sphere are harmonic densities of low degree. Among these are the 2-parameter von-Mises distribution on the circle and the 5-parameter Kent distribution on the sphere. These are both exponential families, about which Mardia & Jupp (1999) write (p. 175): "Although exponential models have many pleasant inferential properties, the need to evaluate the normalizing constant (or at least the first derivative of its logarithm) can be a practical difficulty."

What is a practical difficulty for the unimodal distributions with few parameters mentioned above becomes a showstopper for more flexible exponential families with many parameters. The harmonic exponential family for the circle is known as the generalized von-Mises distribution, and can be defined for any band-limit / order / degree L. However, no scalable maximum-likelihood estimation algorithm is known. Gatto & Jammalamadaka (2007) (who work only with the 4-parameter $L = 2$ distribution) compute the normalizing constant by a truncated infinite sum, where each term involves an expensive Bessel function evaluation.

Beran (1979) studies an exponential family that is equivalent to the harmonic exponential family on the sphere, but does not provide a scalable learning algorithm. At each iteration, the proposed algorithm computes all $O(L^2)$ moments of the distribution by numerical integration. This requires $O(L^2)$ samples per integration, making the per-iteration cost $O(L^4)$. Beran further suggests using a second order optimization method, which would further increase the per-iteration cost to $O(L^6)$.

This is clearly not feasible when L is measured in the hundreds, and parameter counts in the 10's of thousands, as is needed in the experiments reported in section 6. The algorithm described in this paper is simple, generic across manifolds, and fast (per-iteration complexity $O(L^2 \log^2 L)$ in the spherical case, for instance). It can be applied to any manifold for which an FFT has been developed.

3. Preliminaries

The manifolds we consider in this paper are either Lie groups or closely related manifolds called homogeneous spaces, and the sufficient statistics that we use come from harmonic analysis on these manifolds. Since these concepts are not widely known in machine learning, we will review them in this section. For more details, we refer the reader to Chirikjian & Kyatkin (2001); Sugiura (1990); Kondor (2008); Goodman & Wallach (2009).

3.1. Lie Groups

A *transformation group* G is a set of invertible transformations that is closed under composition and taking inverses: for any $g, h \in G$, the composition gh is again a member of G, and so are the inverses g^{-1} and h^{-1}.

A *Lie group* is a group that is also a differentiable manifold. For example, the group $\mathrm{SO}(3)$ of 3D rotations is a Lie group. It can be represented as a set of matrices,

$$\mathrm{SO}(3) = \{R \in \mathbb{R}^{3\times 3} \mid RR^T = I,\, \det(R) = 1\}, \quad (1)$$

so one way to think about $\mathrm{SO}(3)$ is as a 3-dimensional manifold embedded in $\mathbb{R}^{3\times 3}$. For technical reasons, we will further restrict our attention to *compact Lie groups*, i.e. Lie groups that are closed and bounded.

3.2. Harmonic analysis on compact Lie groups

The basic idea of the generalized Fourier transform on compact Lie groups is to expand a function $f : G \to \mathbb{C}$ as a linear combination of carefully chosen basis functions. These basis functions have very special properties, because they are the matrix elements of *irreducible unitary representations* (IURs) of G. These terms will now be explained.

A *representation* of a group G on a vector space V is a map R from the group to the set of invertible linear transformations of V that preserves the group structure in the following sense:

$$R(gh) = R(g)R(h). \tag{2}$$

Note that gh denotes composition of group elements while $R(g)R(h)$ denotes matrix multiplication (once we choose a basis for V, that is).

In computer vision, we encounter group representations in the following way. An image is represented as a vector $x \in \mathbb{R}^n$ of pixel intensities, and to be concrete, we take G to be the group $\mathrm{SO}(2)$ consisting of rotations of the plane. Then $R(\theta)$ is the matrix such that $R(\theta)x$ is the image x rotated by angle θ.

As this example shows, many representations do not change the norm of the vectors on which they act: $\forall g \in G, \forall x \in V : \|U(g)\,x\| = \|x\|$. Such representations are called *unitary*. Unitary representations tend to be easier to work with both analytically and numerically.

Given a unitary representation U and a unitary matrix F, one can define an equivalent representation $T(g) = F^{-1}U(g)F$. In computer vision one can think of F as a matrix containing image features in the rows. One can now try to find F such that for every g, the matrix $F^{-1}U(g)F$ is block diagonal with the same block structure. If we continue to block-diagonalize until no further diagonalization is possible, we end up with blocks called *irreducible unitary representations*[1]. The IURs of a compact group can be indexed by a discrete index λ, and we denote the IURs as $U^\lambda(g)$.

As an example, consider the 2D rotation group $\mathrm{SO}(2)$. Since $\mathrm{SO}(2)$ is commutative, its representation matrices can be jointly diagonalized and so the IURs of $\mathrm{SO}(2)$ are 1×1 matrices:

$$U^\lambda_{00}(g) = e^{i\lambda g}. \tag{3}$$

They satisfy the composition rule $e^{i\lambda(g+h)} = e^{i\lambda g}e^{i\lambda h}$, which is the manifestation of eq. 2 for this representation. The standard Fourier series of a function on the circle is an expansion in terms of these matrix elements, which shows that standard Fourier analysis is a special case of the more general transform to be defined shortly.

The matrix elements of IURs of $\mathrm{SO}(3)$ are known as Wigner D-functions. They are defined for $\lambda = 0, 1, 2, \ldots$ and $-\lambda \le m, n \le \lambda$, so the irreducible representations are $(2\lambda + 1)$-dimensional. Wigner D-functions can be expressed as sums over products of sinusoids or complex exponentials (Pinchon & Hoggan, 2007), but the formulae are somewhat unwieldy so that it is easier to think only about their general properties.

The most important general property of the matrix elements of IURs is that they are orthogonal:

$$\begin{aligned}\langle U^\lambda_{mn}(g),\, U^{\lambda'}_{m'n'}(g)\rangle &\equiv \int_G U^\lambda_{mn}(g)\,\overline{U^{\lambda'}_{m'n'}(g)}\,d\mu(g) \\ &= \frac{\delta_{\lambda\lambda'}\delta_{mm'}\delta_{nn'}}{\dim\lambda}.\end{aligned} \tag{4}$$

Here μ is the normalized Haar measure, which is the natural way to measure volumes in G (Sugiura, 1990), and $\dim\lambda$ is the dimension of the representation. One can verify that the complex exponentials $e^{i\lambda g}$ are indeed orthonormal with $d\mu(g) = \frac{dg}{2\pi}$ and $\dim\lambda = 1$.

Intuitively, the matrix elements are like a "complete orthogonal basis" for the space $L^2(G)$ of square integrable functions on G. That is, it can be proven that any function $f \in L^2(G)$ can be written as

$$f(g) = \sum_\lambda \sum_{mn} \eta^\lambda_{mn} T^\lambda_{mn}(g) \equiv \left[\mathcal{F}^{-1}\eta\right](g), \tag{5}$$

where $T^\lambda_{mn}(g) = \sqrt{\dim\lambda}\,U^\lambda_{mn}(g)$ are the L^2-normalized matrix elements and η are the Fourier coefficients of f.

Integrating eq. 5 against a matrix element and using orthonormality, we find:

$$\eta^\lambda_{mn} = \int_G f(g)\overline{T^\lambda_{mn}(g)}d\mu(g). \tag{6}$$

This is the (generalized) Fourier transform for compact groups, denoted

$$\eta = \mathcal{F}f.$$

Fast and exact algorithms for the computation of Fourier coefficients from samples of bandlimited functions on the rotation groups $\mathrm{SO}(2)$ and $\mathrm{SO}(3)$ have been developed, and the theory required to construct such algorithms for general compact Lie groups is understood (Maslen & Rockmore, 1997). The group $\mathrm{SO}(2)$ is isomorphic to the circle, so for $G = \mathrm{SO}(2)$ equation 5 reduces to a standard Fourier series on the circle, for which the well-known $O(L \log L)$ FFT algorithm can be used. The $\mathrm{SO}(3)$ FFT has complexity $O(L^3 \log^2 L)$ for bandlimit L (Maslen & Rockmore, 1997; Kostelec & Rockmore, 2008; Potts et al.,

[1] Technically, we have defined the slightly easier to understand notion of *indecomposability*, which in this context implies irreducibility.

2009). This is a tremendous speedup compared to the naive $O(L^6)$ algorithm, and the algorithms presented in this paper would certainly not be feasible without the FFT.

In section 4 we discuss how these generalized FFT algorithms can be used to efficiently compute moments, but first we discuss the generalization of the Fourier transform on compact Lie groups to the Fourier transform on certain manifolds that are not groups.

3.3. Harmonic analysis on homogeneous spaces

A *homogeneous space* for a Lie group G is a manifold H such that for any two points $x, y \in H$ we can find a transformation $g \in G$ with $gx = y$. For example, the plane is a homogeneous space for the translation group, and the sphere is a homogeneous space for the 3D rotation group. The plane is not a homogeneous space for the 2D rotation group, because points at different radii cannot be rotated into each other.

If we pick an origin $o \in H$, such as the north pole of the sphere, we can identify any other point $h \in H$ by specifying how to transform the origin to get there: $h = go$. This identification will not be unique, though, if there is a nontrivial subgroup K of G containing transformations that leave the origin invariant: $K = \{k \in G \,|\, ko = o\}$. This is because if $h = go$, then also $h = gko$, so both g and gk identify h. On the sphere for example, we can transform the north-pole into a point h by first doing an arbitrary rotation around the north-pole axis (which leaves the north-pole unchanged) and then rotating the result to h.

Hence, one can think of the *points* $h = go$ in a homogeneous space H as *sets* $gK = \{gk \,|\, k \in K\}$ (called cosets) of group elements that are equivalent with respect to their effect on an arbitrarily chosen origin o of H. It follows that one can think of functions on a homogeneous space as functions on the group, with the special property that they are (right) invariant to transformations from K:

$$f(gk) = f(g) \quad \forall g \in G, k \in K, \tag{7}$$

because right-multiplication by k will only shuffle the elements within each coset. Finally, one can show (see supplementary material) that in a suitable basis, a subset of the matrix elements of IURs form a basis for the linear space of square-integrable functions on G with this invariance property, which allows us to define the Fourier transform also for functions on H.

The exact same equations (6 and 5) that define the Fourier and inverse Fourier transform for a compact Lie group, define these transforms for a compact homogeneous space, but only a subset of the coefficients η^λ_{mn} will be non-zero. For the sphere, the matrix elements T^λ_{m0} are equal[2] to the spherical harmonics Y^λ_m, which form a basis for $L^2(S^2)$. Fast spherical Fourier transform algorithms were developed by Driscoll & Healy (1994).

3.4. Exponential families

An exponential family is a class of densities of the form:

$$p(g \,|\, \eta) = \frac{1}{Z_\eta} \exp\left(\eta \cdot T(g)\right). \tag{8}$$

It is determined by a choice of *sufficient statistics* T, that take the random variable g and produce a vector of real statistics $T(g)$.

To learn the parameters η, one can perform gradient-based optimization of the log-likelihood of a set of iid samples $g_1, \ldots, g_N$. The gradient is the moment discrepancy:

$$\nabla_\eta \left(\frac{1}{N} \sum_{i=1}^{N} \ln p(g_i \,|\, \eta) \right) = \bar{T} - \mathbb{E}_{p(g|\eta)}[T(g)], \tag{9}$$

where $\bar{T} = \frac{1}{N}\sum_{i=1}^{N} T(g_i)$ are the empirical moments. There is generally no closed form for the analytical moments (the expectation in eq. 9), so a numerical approximation is needed.

4. Harmonic Exponential Families

We define a harmonic exponential family on a group or homogeneous space as an exponential family where the sufficient statistics are given by a finite number of matrix elements of IURs. This makes sense only if the function $\eta \cdot T(g)$ is real-valued, so that it can be interpreted as an unnormalized log-probability. The easiest way to guarantee this is to take η to be real, and to use real functions obtained as a sparse linear combination of complex matrix elements $T(g)$ as sufficient statistics. For example, $\frac{1}{2}(e^{i\lambda\theta} + e^{-i\lambda\theta}) = \cos(\lambda\theta)$. From here on, we take T to be real, L^2-normalized functions and $\mathcal{F}$ the expansion of a real function in terms of real basis functions T.

The key observation that leads to an efficient algorithm for computing the moments of a harmonic density is that the moments of such a density are its Fourier coefficients:

$$\mathbb{E}_{p(g|\eta)}\left[T(g)\right] = \int_G p(g|\eta)\, T(g)\, d\mu(g) = \mathcal{F}\, p. \tag{10}$$

Hence, one can obtain all J moments at once by sampling p on a finite grid and then computing its Fourier transform using a fast algorithm. As explained in section 4.1, the discretization error can be made extremely small using only $O(J)$ spatial samples.

[2] Various normalization and phase conventions are in use for the spherical harmonics, but it is enough to know that $Y^\lambda_m \propto T^\lambda_{m0}$.

However, even evaluating p at a single position takes $O(J)$ computations when using J sufficient statistics so that the overall complexity is still $O(J^2)$. Furthermore, in order to evaluate p we need to know the normalizing constant Z_η.

The following derivation shows that we can work with the unnormalized density $\varphi(g \,|\, \eta) = \exp(\eta \cdot T(g))$ instead:

$$\begin{aligned}[\mathcal{F}p]^\lambda_{mn} &= \int_G p(g \,|\, \eta)\, T^\lambda_{mn}(g)\, d\mu(g) \\ &= \frac{1}{Z_\eta} \int_G \varphi(g \,|\, \eta)\, T^\lambda_{mn}(g)\, d\mu(g) \\ &= \frac{[\mathcal{F}\varphi]^\lambda_{mn}}{[\mathcal{F}\varphi]^0_{00}},\end{aligned} \tag{11}$$

The last step uses the fact that $T^0_{00}(g) = 1$ so that $[\mathcal{F}\varphi]^0_{00}$ is equal to the normalizing constant:

$$[\mathcal{F}\varphi]^0_{00} = \int_G \varphi(g \,|\, \eta)\, T^0_{00}(g)\, d\mu(g) = Z_\eta \tag{12}$$

Next, observe that we can evaluate $\ln \varphi$ efficiently at $O(J)$ spatial points using the inverse FFT:

$$\ln \varphi(g \,|\, \eta) = \eta \cdot T(g) = \left[\mathcal{F}^{-1}\eta\right](g) \tag{13}$$

This computation is exact, because the log-density is bandlimited (i.e. there are only finitely many parameters). Element-wise exponentiation then gives us φ evaluated on a grid.

So we have an efficient algorithm for computing moments:

1. Compute $\varphi = \exp(\mathcal{F}^{-1}\eta)$.
2. Compute $M = \mathcal{F}\,\varphi$
3. Compute $\mathbb{E}_{p(g|\eta)}[T(g)] = M/M^0_{00}$.

To make this computation numerically stable for highly peaked densities, one should apply the "log-Fourier-exp" trick described in the supplementary material.

4.1. Approximation quality

Even though the Fourier coefficients are defined as definite integrals (eq. 6), the discrete FFT algorithms compute *exact* Fourier coefficients, provided the function from which the discrete samples were gathered is *bandlimited*. A function is bandlimited if the coefficients η^λ_{mn} are zero for λ greater than the band-limit L. Although the function $\ln \varphi(g) = \eta \cdot T(g)$ is bandlimited, the function $\varphi(g) = \exp(\eta \cdot T(g))$ is not, so the computed coefficients are not exactly equal to the Fourier coefficients of φ.

However, the function $\varphi(g)$ is smooth (infinitely differentiable), and a standard result in Fourier analysis shows that the spectrum of a smooth function decays to zero asymptotically faster than $O(1/\lambda^n)$ for any n. So our function will be "effectively bandlimited", in the sense that coefficients for λ greater than some pseudo-bandlimit will have negligible values. If L is the maximum degree of the sufficient statistics (the bandlimit of $\eta \cdot T(g)$), one can obtain near-exact moments by computing the FFT up to the pseudo-bandlimit αL for some oversampling factor α. In practice, we use values for α ranging from 2 to 5.

5. Harmonic Densities as Conjugate Priors

In this section we discuss the Bayesian transformation inference problem, where the goal is to infer a posterior over a Lie group of transformations given only a set of correspondence pairs (such as images before and after a camera motion). It turns out that the harmonic exponential families are the conjugate priors for this problem, and again, the generalized FFT is key to performing efficient inference.

The observed data in the Bayesian transformation inference problem are pairs of vectors $(x,\, y)$ in $\mathbb{R}^D$ that could represent images, space-time blocks of video, point-clouds, optical-flow fields, fitted geometric primitives, parameters of a function or other objects. In order to infer anything about a latent transformation g, we must know the group representation $R(g)$ that acts on the observed data. If our data is an image $x : \mathbb{R}^2 \to \mathbb{R}$, we get a representation on the Hilbert space in which x lives: $[R(g)x](p) = x(g^{-1}p)$, where p is a point in the plane. In the finite-dimensional analogue, where x is a vector of pixel intensities, $R(g)$ will be close to a permutation matrix that takes each pixel to its proper new position. For compact groups[3] this representation is unitary, and this is what we will assume for $R(g)$ from now on. If the representation is not known in advance, it can also be learned from data (Cohen & Welling, 2014).

If we assume that observation x is the g-transformed version of y with some independent Gaussian noise with variance σ^2 added, the likelihood function is given by

$$p(x \,|\, y, g) = \mathcal{N}(x \,|\, R(g)y, \sigma^2). \tag{14}$$

As discussed in section 3.2, we can bring R in block-diagonal form by a unitary change of basis: $U(g) = F^{-1}R(g)F$. The matrix U is block diagonal, and the blocks U^λ are equal to the L^2-normalized sufficient statistics T^λ up to a scale factor: $T^\lambda(g) = \sqrt{\dim \lambda}\, U^\lambda(g)$. To simplify the computations, we shall work with data in this new basis: $\hat{x} = Fx$ so that $p(\hat{x} \,|\, \hat{y},\, g) = \mathcal{N}(\hat{x} \,|\, U(g)\hat{y}, \sigma^2)$.

If we now choose as prior $p(g)$ a member of the harmonic exponential family on G, the posterior $p(g \,|\, \hat{x},\, \hat{y})$ is of the

[3] This more generally true for unimodular groups.

same form as the prior:

$$\begin{aligned} p(g \mid \hat{x}, \hat{y}) &\propto p(\hat{x} \mid \hat{y}, g)p(g) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|\hat{x} - U(g)\hat{y}\|^2 + \eta \cdot T(g)\right) \\ &\propto \exp\left(\frac{1}{\sigma^2}\hat{x}^T U(g)\hat{y} + \eta \cdot T(g)\right) \\ &= \exp\left(\sum_\lambda \left(\eta^\lambda + \frac{\hat{x}_\lambda \hat{y}_\lambda^T}{\sigma^2\sqrt{\dim \lambda}}\right) \cdot T^\lambda(g)\right) \end{aligned} \tag{15}$$

i.e. we have a conjugate prior. The derivation relies on the unitarity of the representation: in expanding $\|\hat{x} - U(g)\hat{y}\|^2$, we find a term $\|U(g)\hat{y}\|^2$ which is equal to $\|\hat{y}\|^2$, making the dependence on $U(g)$ *linear* (as it is in the prior).

5.1. Example: Bayesian analysis of camera rotation

To make matters concrete, we show how to compute a posterior over the rotation group $\mathrm{SO}(3)$ given two images taken before and after a camera rotation. An image is modeled as a function $x : S^2 \to \mathbb{R}$ on the sphere, so that a camera rotation $g \in \mathrm{SO}(3)$ acts by rotating this function over the sphere: $[R(g)x](p) = x(g^{-1}p)$. We parameterize points $p \in S^2$ as $p = (\varphi, \theta)$ for $\varphi \in [0, 2\pi]$ and $\theta \in [0, \pi]$. Recall from section 3.3 that we can represent $p \in S^2$ by a coset representative $g_p \in \mathrm{SO}(3)$, which we parameterize using Euler angles as $g_p = (\varphi, \theta, 0)$. The transformation g_p takes the origin of the sphere to p.

It is well known[4] that in this context the matrix F — defined in the previous section as the matrix that block-diagonalizes the representation R — is given by the spherical Fourier transform $\mathcal{F}$, which can be computed by an FFT. This means that if we represent x by its Fourier coefficients $\hat{x} = \mathcal{F}x$, the coefficients of the rotated function $R(g)x(p) = x(g^{-1}p)$ are given by $U(g)\hat{x}$, where $U(g)$ is a block-diagonal matrix with irreducible representations U^λ as blocks. As derived in the previous section, the parameters of the posterior can easily be obtained in this basis by a block-wise outer product.

Figure 1 shows the posterior $p(g, \mid x, y)$ for two synthetic spherical images x^1 and x^2, and their rotations $y^1 = x^1$ (no rotation) and $y^2 = U(0, \pi/3, \pi/2)x^2$. The posterior is plotted as a 3D isocontour in the ZYZ-Euler angle parameter space $(\alpha, \beta, \gamma) \in [0, 2\pi] \times [0, \pi] \times [0, 2\pi]$. Although this plot looks like a box, the actual manifold has the topology of a projective 3-sphere. By construction, the density is spread through the parameter space in a way that is consistent with this topology: no discontinuities arise where wraparound in the parameter space occurs. This is desirable, because the wraparound is not an intrinsic property of the manifold.

[4]The derivation can be found in the supplementary material.

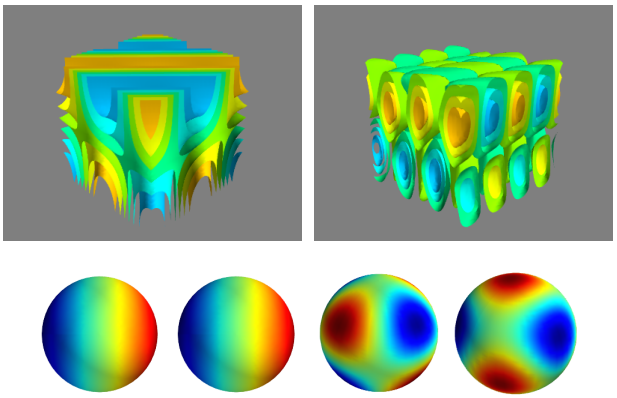


Figure 1. Posterior distributions (top) for two correspondence pairs (bottom).

Due to the symmetry of these figures certain rotations cannot be distinguished from the "true" rotation, so that the modes of the posterior distributions are supported on an entire subgroup of $\mathrm{SO}(3)$. Real images will not have such a high degree of symmetry, but it will nevertheless often be the case that a unique optimal transformation does not exist (Ma et al., 1999). Indeed, current keypoint based transformation estimation methods can easily get confused by repeating structures in an image, such as several identical windows on a building. Although more experimental work is needed, our method has the theoretical advantage that besides keypoints (which form a group representation), it can make use of parts of the image that do not allow for keypoints to be reliably placed (such as edges), while always providing a truthful impression of the degree to which a unique transformation or subgroup can be identified form the data.

5.2. MAP inference

As Bayesians we are done here, but for some applications one may wish to find a single most likely transformation. To find the optimal transformation, first perform posterior inference (steps 1 and 2) and then maximize (step 3):

1. Compute $\hat{x} = \mathcal{F}\,x$ and $\hat{y} = \mathcal{F}\,y$.
2. Compute $\bar{\eta}^\lambda = \eta^\lambda + \frac{1}{\sigma^2\sqrt{\dim \lambda}}\hat{x}_\lambda \hat{y}_\lambda^T$
3. Compute $g^* = \arg\max_i [\mathcal{F}^{-1}\bar{\eta}](g_i)$

The $\arg\max$ ranges over all the points in a finite grid on G used by the FFT synthesis $\mathcal{F}^{-1}$. Optionally, one can refine the optimum g^* by performing a few steps of gradient-based optimization on $\bar{\eta} \cdot T(g)$ to get sub-pixel accuracy.

6. Experiments

6.1. Modelling the spatial distribution of earthquakes

We compare our model and MLE algorithm to a Kent Mixture Model (KMM) on the problem of modelling the spatial distribution of significant earthquakes on the surface of the earth.

We obtained the Significant Earthquake Dataset (NGDC, 2015) from the National Geophysical Datacenter of the National Oceanographic and Athmospheric Administration. In total, the dataset contains 5780 earthquakes with complete information on the position of their epicenter, and 53 earthquakes whose coordinates are missing (these were discarded in our experiments). We did not model the severity of the earthquake, but only the occurrence of significant earthquakes (as defined by (NGDC, 2015)).

6.1.1. Mixture of Kent distributions

The 5-parameter Kent distribution (Mardia & Jupp, 1999) is the spherical analogue of the normal distribution with unconstrained covariance. Being unimodal, the Kent distribution is not flexible enough to describe complicated distributions such as the spatial distribution of earthquakes. The most flexible distribution on the sphere that we have found in the literature is the Kent Mixture Model, first described by Peel et al. (2001). The KMM is trained using the EM algorithm. We use the open source Python implementation of the EM algorithm for KMMs by Höfer (2014).

Unlike the harmonic densities, the log-likelihood of this model is not convex and contains many singularities where a mixture component concentrates on a single data point and decreases its variance indefinitely. For this reason, we perform randomly initialized restarts until the algorithm has found 10 non-degenerate solutions, of which we retain the one with the best cross-validation log-likelihood. No regularization was used, because for the models that could be trained within a reasonable amount of time, no overfitting was observed.

6.1.2. The S^2 harmonic density

The harmonic density on the 2-sphere uses spherical harmonics as sufficient statistics. The empirical moments are easily computed using standard spherical harmonic routines, but we found that for high orders the SciPy routines are slow and numerically unstable. The supplementary material describes a simple, fast, and stable method for the evaluation of spherical harmonics. The computation of spherical harmonics up to band-limit $L = 200$ (for a total of $(L+1)^2 = 40401$ spherical harmonics) for 5780 points on the sphere took half a minute using this method and is performed only once for a given dataset.

For regularization we use a diagonal Gaussian prior on η, where the precision β_m^λ corresponding to the coefficient of Y_m^λ is given by $\beta_m^\lambda = \alpha \dim \lambda = \alpha(2\lambda + 1)$ (for some regularization parameter α). This scheme is inspired by the fact that $\dim \lambda$ is the discrete Plancherel measure (Sugiura, 1990), and the empirical observation that the fitted coefficients become approximately uniform when weighted as $\eta_m^\lambda \sqrt{\dim \lambda}$. Note that adding regularization does not change the convexity of the objective function.

To find maximum a posteriori parameters $\hat{\eta}$ for the spherical harmonic density, we perform iterative gradient-based optimization on the log-posterior. The gradients (moment discrepancies) are computed using the FFT-based method described in section 4. We use the spherical FFT algorithm implemented in the NFFT library (Keiner et al., 2009; Kunis & Potts, 2003). The gradients are fed to a standard implementation of the L-BFGS algorithm.

6.1.3. Results

Figure 2 shows the average train and test log-likelihood over 5 cross-validation folds, for the spherical harmonic density and the mixture of Kent distribution. The plotted values correspond to the regularization settings that yielded the best test log-likelihood.

The KMM reached an average test log-likelihood of -0.37 (with standard deviation of 0.03 over 5 cross-validation folds) using 70 mixture components ($5 \times 70 + 69 = 419$ parameters). The harmonic density reached an average test log-likelihood of $+0.3$ (with standard deviation of 0.036 over 5 cross-validation folds), using bandlimit $L = 140$ (19880 parameters). The HD still outperforms the KMM when given a similar number of parameters: for $L = 20$ ($440 \approx 419$ parameters), the log-likelihood is $-0.28 > -0.37$, with standard deviation 0.028, and for $L = 19$ (399 parameters), the log-likelihood is -0.3 with standard deviation 0.03. The dataset and the learned densities are plotted in figure 3, clearly showing the superiority of the harmonic density.

7. Discussion and Future Work

What could explain the difference in log-likelihood between our model and the KMM? We believe two interrelated factors are driving this difference: the expressiveness of the model and the ease of optimization. Leaving technicalities aside, it is clear that both the spherical harmonic exponential family and the KMM can approximate any well-behaved density, given enough parameters and an optimization oracle. However, as is clear from figure 2, training time becomes prohibitive for the KMM for more than 70 mixture components / 419 parameters, while the harmonic density can efficiently be fit for tens of thousands

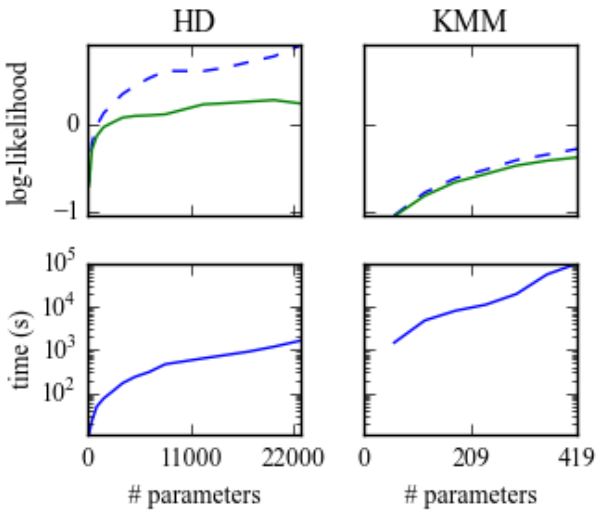
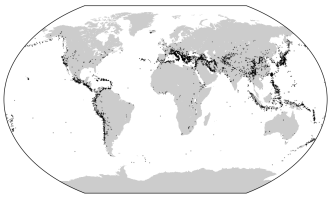
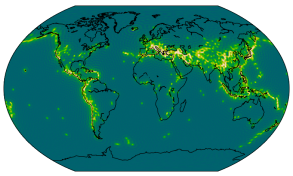


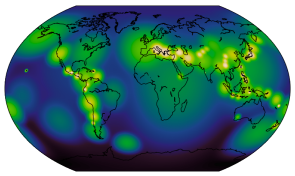
Figure 2. Top: cross-validation train (dashed line) and test (solid line) log-likelihood for the Harmonic Density (HD) and Kent Mixture Model (KMM). Bottom: number of parameters versus training time for both models.



(a) Significant Earthquake Dataset



(b) Harmonic Density



(c) Kent Mixture Model

Figure 3. Log-probability for harmonic density and Kent mixture model, plotted using a perceptually accurate linear-lightness colormap on the same intensity scale.

of parameters.

Furthermore, the KMM training algorithm (EM) can easily get stuck in local optima, or converge on a degenerate solution. This is the main reason for the poor runtime performance; while the KMM code could be further optimized, it is the fact that so many restarts are required to find a good fit that makes the algorithm slow. The log-likelihood function of the harmonic density, on the other hand, is convex, and the L-BFGS optimizer will typically converge to the global optimum in some $20 - 100$ iterations.

As can be seen in figure 3, the harmonic density produces slight ringing artifacts that can be seen only in a log-plot such as this. These are the result of the limited bandwidth of the log-density, and will become progressively less pronounced as the number of parameters is increased. While they are clearly visible in log-space, the actual difference between peaks and valleys is on the order of 10^{-3} for bandlimit $L = 100$. The artifacts are not visible on a non-logarithmic plot (and in such a plot the KMM density is hardly visible at all when plotted on the same intensity scale as the harmonic density, because the peaks are much lower). The harmonic density also tends to prefer heavier tails, which is probably accurate for many problems.

An interesting direction for future work is the extension to non-compact groups. While the mathematical theory becomes much more technical for such groups, (Kyatkin & Chirikjian, 2000) have already succeeded in developing FFT algorithms for the Euclidean motion group which is non-compact. From there it should be relatively straightforward to develop harmonic densities on the Euclidean group. Harmonic densities on the Euclidean, affine or even projective group should find many applications in robotics and computer vision (see e.g. (Kyatkin & Chirikjian, 1999)).

8. Conclusion

We have studied a class of exponential families on compact Lie groups and homogeneous manifolds, which we call harmonic exponential families. We have shown that for these families, maximum likelihood inference, posterior inference and mode seeking can be implemented using very efficient generalized Fast Fourier Transform algorithms. In the Bayesian setting, we have shown that harmonic exponential families appear naturally as conjugate priors in the generic transformation inference problem. Our experiments show that training harmonic densities is fast even for very large numbers of parameters, and that far superior likelihood can be achieved using these models.

Acknowledgements

This research was supported by NWO (grant number NAI.14.108), Facebook and Google.

References

Beran, Rudolf. Exponential Models for Directional Data. *The Annals of Statistics*, 1979.

Chirikjian, G.S. and Kyatkin, A.B. *Engineering Applications of Noncommutative Harmonic Analysis*. CRC

Press, 1 edition, May 2001. ISBN 9781420041767.

Cohen, T. and Welling, M. Learning the Irreducible Representations of Commutative Lie Groups. In *International Conference on Machine Learning (ICML)*, volume 32, 2014.

Diaconis, P. *Group representations in probability and statistics*. Hayward, CA: Institute of Mathematical Statistics, 1988. ISBN 0940600145.

Driscoll, JR and Healy, DM. Computing Fourier transforms and convolutions on the 2-sphere. *Advances in applied mathematics*, 1994.

Gatto, Riccardo and Jammalamadaka, Sreenivasa Rao. The generalized von Mises distribution. *Statistical Methodology*, 4(3):341–353, 2007. ISSN 15723127.

Goodman, Roe and Wallach, Nolan R. *Symmetry, Representations, and Invariants*. Springer, 2009. ISBN 978-0-387-79852-3.

Höfer, Sebastian. Kent mixture model for scikit-learn, 2014. URL https://github.com/shoefer/scikit-learn.

Keiner, J., Kunis, S., and Potts, D. Using NFFT 3 - a software library for various nonequispaced fast Fourier transforms. *ACM Transactions on Mathematical Software*, 36:1–30, 2009.

Kondor, Risi. *Group theoretical methods in machine learning*. PhD thesis, Columbia University, 2008.

Kostelec, Peter J. and Rockmore, Daniel N. FFTs on the rotation group. *Journal of Fourier Analysis and Applications*, 14(2):145–179, 2008.

Kunis, Stefan and Potts, Daniel. Fast spherical Fourier algorithms. *Journal of Computational and Applied Mathematics*, 161:75–98, 2003.

Kyatkin, AB and Chirikjian, GS. Computation of Robot Configuration and Workspaces via the Fourier Transform on the Discrete-Motion Group. *The International Journal of Robotics Research*, 1999.

Kyatkin, Alexander B and Chirikjian, Gregory S. Algorithms for Fast Convolutions on Motion Groups. *Applied and Computational Harmonic Analysis*, 9(2):220–241, September 2000.

Ma, Y, Soatto, S, Kosecka, J, Sastry, S, and Košecká, J. Euclidean reconstruction and reprojection up to subgroups. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, volume 2, pp. 773—-780. California Univ., Berkeley, CA, IEEE, 1999.

Mardia, K. V. and Jupp, P. E. *Directional Statistics*. John Wiley & Sons, 1 edition, 1999. ISBN 0471953334.

Maslen, DK and Rockmore, DN. Generalized FFTs – a survey of some recent results. *Groups and Computation II*, 1997.

NGDC. National Geophysical Data Center / World Data Service: Significant Earthquake Database. [accessed 01-18-2015], 2015. URL http://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=1&d=1.

Peel, David, Whiten, William J, and McLachlan, Geoffrey J. Fitting Mixtures of Kent Distributions to Aid in Joint Set Identification. *Journal of the American Statistical Association*, 96(453):56–63, 2001. ISSN 0162-1459. doi: 10.1198/016214501750332974.

Pinchon, Didier and Hoggan, Philip E. Rotation matrices for real spherical harmonics: general rotations of atomic orbitals in space-fixed axes. *Journal of Physics A: Mathematical and Theoretical*, 40(7):1597–1610, February 2007. ISSN 1751-8113.

Potts, Daniel, Prestin, J, and Vollrath, A. A fast algorithm for nonequispaced Fourier transforms on the rotation group. *Numerical Algorithms*, pp. 1–28, 2009.

Sugiura, Mitsuo. *Unitary Representations and Harmonic Analysis*. John Wiley & Sons, New York, London, Sydney, Toronto, 2nd edition, 1990.

Wainwright, Martin J. and Jordan, Michael I. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2007. ISSN 1935-8237.