entropy

MDPI

Review

Approximation of Densities on Riemannian Manifolds

Alice le Brigant [†] **and Stéphane Puechmorel** [*,†]

Ecole Nationale de l'Aviation Civile, Université de Toulouse, 31055 Toulouse, France; alice.le-brigant@enac.fr
* Correspondence: stephane.puechmorel@enac.fr; Tel.: +33-5-62259503
† These authors contributed equally to this work.

Received: 11 December 2018; Accepted: 3 January 2019; Published: 9 January 2019

Abstract: Finding an approximate probability distribution best representing a sample on a measure space is one of the most basic operations in statistics. Many procedures were designed for that purpose when the underlying space is a finite dimensional Euclidean space. In applications, however, such a simple setting may not be adapted and one has to consider data living on a Riemannian manifold. The lack of unique generalizations of the classical distributions, along with theoretical and numerical obstructions require several options to be considered. The present work surveys some possible extensions of well known families of densities to the Riemannian setting, both for parametric and non-parametric estimation.

Keywords: quantization; directional densities; exponential family; group invariance; Riemannian manifold

1. Introduction

In probability and statistics, random variables whose law admits a probability density (with respect to e.g., the Lebesgue measure) are more tractable than general ones, both from the theoretical and the algorithmic point of view. When dealing with experimental data, the density is generally unknown and must be estimated.

In many cases, it is a function belonging to a given family, defined on the image of the random variable. When the family depends on a finite number of parameters, the estimation problem boils down to finding a point in the parameter space minimizing a goodness-of-fit criterion. Methods pertaining to this class are referred to as parametric. On the other hand, when prior information on the true density is lacking or when the parametric approach is too complicated or expensive from a computational point of view, it may be more pertinent to use another class of methods, the *non-parametric* ones, that does not rely on fitting parameters. In this case, the true density is computed from the samples themselves, often by summing copies of a model density known as a kernel function.

On the other hand, some applications require an approximation rather than a proper estimation of the density of a given dataset. When dealing with large datasets, it can be interesting to search for a summary of the empirical distribution in the form of a discrete probability measure with a given number of supporting points. This is known as the quantization problem and has received much attention from the signal processing community. It is worth noticing that finding an optimal quantization is closely related to clustering: each supporting point in the discrete distribution can be thought as a cluster center, with a membership function that associates with a sample the closest supporting point.

In all the above cases, density approximation is central. When the considered random variables are defined on a finite dimensional real vector space, the problem has been extensively studied [1,2]. However, in many applications, the data are best modeled as elements of a Riemannian manifold, and approximation procedures have to be adapted. Classical examples include the sphere $\mathbb{S}^d$, the d-dimensional torus and some matrix spaces. In medical imaging, for example, diffusion tensor images have pixels taking their values in the cone of symmetric positive definite matrices. The same type of data arises when assessing a complexity level to a traffic situation in the air transportation system context. The geometry of correlation matrices is related to the hyperbolic space and is very informative in signal processing applications. An important issue is the lack of a unique extensions of commonly used distributions, like the Gaussian one. The problem is even more acute if one wants to use these model distributions as elementary bricks for approximating more complex ones. This paper will survey the different options offered to estimate and approximate probability densities on Riemannian manifolds. After a brief summary of the main notions of Riemannian geometry that will be needed in the sequel, we present parametric estimation in Section 3. Sections 4 and 5 are devoted to non-parametric estimation, and Section 6 to optimal quantization.

2. Some Notions from Riemannian Geometry

2.1. Differentiable Manifolds

A *topological manifold* of dimension d is a topological space M that can be locally approximated at any point $p \in M$ by a subset of $\mathbb{R}^d$ through a so-called *local chart*, that is, a homeomorphism $\varphi : U \to \mathbb{R}^d$ (a continuous, bijective map with continuous inverse) defined on a neighborhood U of p. A collection of local charts $(U_i, \varphi_i)_{i \in I}$ such that the union of the U_i's covers all of M is called an *atlas*, and M is said to be *differentiable* if one can go from one local chart to another in a differentiable way. More precisely, M is of class C^k if the *transition maps* $\varphi_i \circ \varphi_j^{-1}$ defined by composing chart maps with intersecting domains $U_i \cap U_j \neq \emptyset$ are C^k maps from $\varphi_j(U_i \cap U_j) \subset \mathbb{R}^d$ to $\varphi_i(U_i \cap U_j) \subset \mathbb{R}^d$. Moreover, if the Jacobian determinants of the transition maps are positive, the manifold is said to be *orientable*. This property is mandatory to define a volume form, an object that is used repeatedly in the sequel to define densities. However, in case of non-orientable manifolds, it is still possible to use the weaker notion of Riemannian density. It is an odd-type differential form in the sense of [3]. In the rest of the paper, for the sake of simplicity, we will always consider smooth (C^∞), orientable manifolds.

The local charts allow us to transpose (local) computations to the familiar Euclidean framework and to export definitions from that setting. Given another differentiable manifold N of dimension n, we can naturally define a mapping $F : M \to N$ to be differentiable at a point $p \in M$ if a given (or equivalently, any) representation $\psi \circ F \circ \varphi^{-1}$ of F using local charts $\varphi : U \to \mathbb{R}^d$ of M and $\psi : V \to \mathbb{R}^n$ of N is differentiable at $\phi(p)$ as a map from $\varphi^{-1}(U) \subset \mathbb{R}^d$ to $\mathbb{R}^n$. In what follows, we may place ourselves in a local chart (U, φ) and use the corresponding local coordinates $(x_1(p), \ldots, x_d(p)) := \varphi(p)$.

2.2. Tangent and Cotangent Vectors

A *tangent vector* at a point p can be seen as the intrinsic (i.e., compatible with chart transition functions) velocity of a curve in M passing through p, as well as a derivation acting on the algebra F_p of germs at p of smooth real-valued functions $f : M \to \mathbb{R}$ [4] (Chapter 1.7).

More precisely, let $\alpha : (-\epsilon, \epsilon) \to M$ be a smooth curve on M, and let $p = \alpha(0)$. For any smooth function $f \colon M \to \mathbb{R}$ with germ $[f] \in F_p$, the derivative

$$\left.\frac{d}{dt}\right|_{t=0} f \circ \alpha(t)$$

depends only on $[f]$ and will be denoted as $X([f])$. The operator $X \colon F_p \to \mathbb{R}$ is obviously linear and satisfies the Liebniz rule:

$$X\left([fg]\right) = X([f])[g](p) + [f](p)X([g]).$$

X is called the tangent vector to α at p. Using Hadamard's lemma in a chart, one can show that X can be in fact represented by a vector in $\mathbb{R}^d$, hence the name. Please note that X as a derivation is coordinate-free, while it depends on the local coordinates when represented in $\mathbb{R}^d$.

A tangent vector at p is a tangent vector to a certain curve passing through p at time zero. The set of all tangent vectors at p defines the *tangent space* T_pM, a vector space of same dimension as M, and the collection of all tangent spaces defines the *tangent bundle* $TM = \cup_{p\in M} T_pM$. Given a coordinate chart $\varphi = (x_1, \dots, x_d)$, the tangent vectors defining partial derivation according to coordinates x_i are denoted by $\frac{\partial}{\partial x_1}(p), \dots, \frac{\partial}{\partial x_d}(p)$ and define a basis of T_pM. As any vector space, the tangent space at p admits a dual space T_p^*M called the *cotangent space*, and composed of linear forms $z_p : T_pM \to \mathbb{R}$, also called *cotangent vectors*. The basis of T_pM^* in local coordinates is denoted by $dx_1(p), \dots, dx_d(p)$.

2.3. Pullback and Pushforward

Associated with the dual notions of tangent and cotangent vectors are the dual notions of *pushforward* and *pullback*. Given a smooth map $F : M \to N$ between two smooth manifolds, the pushforward of F is a linear map $F_* : TM \to TN$ that maps tangent vectors X_p at point $p \in M$ to tangent vectors F_*X_p at the image point $F(p) \in N$. The vector F_*X_p can be defined as acting on real-valued functions $f : N \to \mathbb{R}$ or equivalently, as the velocity vector of a curve $\alpha : (-\epsilon, \epsilon) \to M$ passing through p at time zero with speed $\dot{\alpha}(0) = X_p$,

$$(F_*X_p)(f) = X_p(f \circ F), \quad F_*X_p = \left.\frac{d}{dt}\right|_{t=0} F \circ \alpha(t).$$

The pushforward is also called the *differential* of F and can also be denoted $d_pF := (F_*)_p$. Symmetrically, the pullback maps cotangent vectors $z_{F(p)}$ at $F(p) \in N$ to cotangent vectors at $p \in M$, acting on tangent vectors $X_p \in T_pM$ as

$$(F^* z_{F(p)})(X_p) = z_{F(p)}(F_*X_p).$$

2.4. Vector Fields and Covariant Derivatives

A *vector field* is a mapping $X : M \to TM$ that associates with each point p a tangent vector $X_p \in T_pM$. Just as tangent vectors, it acts on differentiable functions $f : M \to \mathbb{R}$ in a way that can be written, in local coordinates, as

$$X_p(f) = \sum_{i=1}^{d} a_i(p) \frac{\partial f}{\partial x_i}(p).$$

It is possible to take the derivative of a vector field with respect to another using an *affine connection*, that is, a functional ∇ that acts on pairs of vector fields $(X, Y) \mapsto \nabla_X Y$ according to the following rules

$$\nabla_{fX+Y} Z = f\nabla_X Z + \nabla_Y Z,$$
$$\nabla_X(fY) = X(f)Y + f\nabla_X Y, \quad \nabla_X(Y + Z) = \nabla_X Y + \nabla_X Z.$$

This action is referred to as *covariant derivative*. Vector fields $V : (-\epsilon, \epsilon) \to TM$ can also be defined along a curve $\alpha(t)$, that is, $V(t)$ is an element of $T_{\alpha(t)}M)$ for all t. Then, covariant derivative is sometimes denoted by $\frac{DV}{dt} := \nabla_{\dot{\alpha}(t)} V$.

2.5. Riemannian Metric and Geodesics

The possibility to compute angles and lengths in a differentiable manifold is given by a *Riemannian metric*, i.e., a smoothly varying inner product $g_p : T_pM \times T_pM \to \mathbb{R}$ defined on each tangent space T_pM at $p \in M$ (recall that T_pM is a vector space). The subscript p in the metric will often be omitted and the associated norm will be denoted by $\| \cdot \| := g(\cdot, \cdot)$. There is only one affine connection that is symmetric, meaning $XY - YX = \nabla_X Y - \nabla_Y X$, and compatible with the Riemannian metric, which is

$$\frac{d}{dt} g(U, V) = g(\tfrac{DU}{dt}, V) + g(U, \tfrac{DV}{dt}),$$

for any vector fields U, V along a curve α. It is called the *Levi–Civita connection* associated with g and will be denoted by ∇ from now on. Just as the Euclidean distance can be measured as the length of a straight line, distances in a Riemannian manifold are computed through the length of minimizing *geodesics*. The geodesics of M are the curves γ satisfying the relation $\nabla_{\dot{\gamma}} \dot{\gamma} = 0$, which implies that their speed has constant norm $\|\dot{\gamma}(t)\| = cst$. They are also the local minimizers of the arc length functional l:

$$l \colon \gamma \mapsto \int_0^1 \|\dot{\gamma}(t)\| dt$$

if curves are assumed, without loss of generality, to be defined over the interval $[0, 1]$. When it exists, the length of the shortest geodesic linking two points defines their geodesic distance. The *cut locus* of $p \in M$ is the set of points where the geodesics starting at p stop being minimizing, and the *injectivity radius* at p is its distance to the cut locus. The global injectivity radius of the manifold is the infimum of the injectivity radii over all points of M.

2.6. Exponential and Logarithm Maps

From the geodesics of M, we can now define the *exponential map* at point p, a diffeomorphism (i.e., a differentiable, bijective map of differentiable inverse) denoted by $\exp_p$, which maps a tangent vector v of an open ball $B(0, r) \subset T_p M$ centered in 0 to the endpoint $\gamma(1) =: \exp_p(v)$ of the geodesic $\gamma : [0, 1] \to M$ verifying $\gamma(0) = p$, $\dot{\gamma}(0) = v$. Intuitively, the exponential map moves the point p along the geodesic starting from p at speed v and stops after covering the length $\|v\|$. Conversely, the inverse of the exponential map $\log_p(q) := \exp_p^{-1}(q)$ gives the vector that maps p to q. The image by the exponential map of the open ball $B(0, r) \subset T_p M$, with r less than the injectivity radius at p, is called the *geodesic ball* of radius $r > 0$ centered in p.

2.7. Curvature and Jacobi Fields

The *curvature tensor* of M associates with any pair of vector fields (X, Y) on M a linear mapping $R(X, Y)$ on the space of vector fields, defined for all vector field Z by

$$R(X, Y)Z := \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z,$$

where $[X, Y] := XY - YX$ denotes the *Lie bracket*. Another way to characterize the curvature of M is through the *sectional curvature*, which is defined for any two-dimensional subspace $\sigma \subset T_p M$ of the tangent space at p by

$$K(\sigma) = \frac{g_p(R(u, v)v, u)}{g_p(u, u) g_p(v, v) - g_p(u, v)^2},$$

if u, v are two linearly independent vectors that span σ. Due to the curvature of a manifold, geodesics spreading out from a given point p can either diverge (negative curvature) or converge (positive curvature). The way these geodesics spread out is described by the *Jacobi fields*. If $\{t \mapsto \exp_p(tv(s)), s \in (-\epsilon, \epsilon)\}$ is a sheave of geodesics starting from the same point p at speeds $v(s) \in T_p M$, and $\gamma(t) := \exp_p(tv(0))$ denotes the one at the center of the sheave, the vector field along γ defined for all t by

$$J(t) = \left.\frac{d}{ds}\right|_{s=0} \exp_p(tv(s))$$

is a Jacobi field along γ. It is the only one with initial conditions $J(0) = v(0)$ and $\frac{DJ}{dt}(0) = \dot{v}(0)$ (where we identify the two vector spaces $T_pM \approx T_{v(0)}T_pM$) verifying the Jacobi equation $\frac{D^2}{dt^2}J + R(J, \dot{\gamma})\dot{\gamma} = 0$.

2.8. Measures and Integration over a Riemannian Manifold

A *differentiable k-form* ω on an orientable d-dimensional manifold M associates with all $p \in M$ an alternating multilinear function $\omega_p : (T_pM)^d \to \mathbb{R}$ (i.e., ω_p associates zero to any d-tuple with a repetition). If ω is a differentiable k-form on a manifold N, then any smooth map $F : M \to N$ induces by pullback a k-form on M acting on k-tuples of tangent vectors at $p \in M$ as

$$(F^*\omega)_p(u_1, \dots, u_k) = \omega_{F(p)}(F_*u_1, \dots, F_*u_k).$$

The *volume forms* of M are the differential forms of maximal degree d (the dimension of the manifold), and are the only ones that can be integrated over M. If (U, φ) is a local chart such that $\operatorname{supp}\omega \subset U$, then $(\varphi^{-1})^*\omega$ is a d-form on $\mathbb{R}^d$, and so it admits a density $f : \varphi^{-1}(U) \mapsto \mathbb{R}$ with respect to the volume element defined in local coordinates by the exterior product $dx_1 \wedge \dots \wedge dx_d$. The integral of the volume form ω is then defined by

$$\int_U \omega := \int_{\varphi^{-1}(U)} (\varphi^{-1})^*\omega = \int_{\varphi^{-1}(U)} f(x)dx.$$

Every volume form defines a measure on M, which is written by extension in local coordinates $d\mu = f dx_1 \wedge \dots \wedge dx_d$ or $d\mu = f dx$ for short. The *Riemannian measure* is the volume form which density is given by the square root of the determinant of the Riemannian metric, i.e.,

$$d\text{vol}(x) = \sqrt{\det G(x)}dx,$$

where $G(x)$ is the $d \times d$ matrix with entries

$$G_{ij}(x) = g_x\left(\frac{\partial}{\partial x_i}(x), \frac{\partial}{\partial x_j}(x)\right).$$

The Riemannian measure will play the role of the Lebesgue measure for integrals defined on M.

2.9. The Laplace–Beltrami Operator

Finally, in order to make this work self-contained, we introduce the generalization of the Laplacian to manifolds, namely, the *Laplace–Beltrami operator*. Let X be a vector field on M and $\phi_X(t, x)$, $(t, x) \in (-\epsilon, \epsilon) \times U$ its *local flow* in a neighborhood U of $p \in M$, i.e., such that for all x, $t \mapsto \phi_X(t, x)$ is the unique curve verifying $\partial_t\phi_X(t, x) = X_{\phi_X(t,x)}$ and $\phi_X(0, x) = x$. Then, the *Lie derivative* of the volume form along the vector field X is given by the derivative of its pullback by the flow of X

$$\mathcal{L}_X\text{vol} = \left.\frac{d}{dt}\right|_{t=0} \phi_X(t, \cdot)^*\text{vol}.$$

Intuitively, it measures the way infinitesimal volume is transported by X. Since $\mathcal{L}_X\text{vol}$ is a d-form, it admits a density with respect to the Riemannian volume form, which is defined to be the *divergence* of X, i.e., $\mathcal{L}_X\text{vol} = (\text{divX})\text{vol}$. Then, the Laplace–Beltrami operator of a function $f : M \to \mathbb{R}$ is, just as the Euclidean case, defined as the divergence of its gradient

$$\Delta f = \text{div}(\text{gradf}),$$

where the gradient is linked to the differential (or pushforward) as $g_p(\text{grad}_\text{p}\text{f}, \text{X}_\text{p}) = \text{d}_\text{p}\text{f}(\text{X}_\text{p})$ for any tangent vector $X_p \in T_pM$.

The Laplace–Beltrami operator can defined alternatively using the Levi–Civita connection. Let X, Y be vector fields and $f : M \to \mathbb{R}$ and $f : M \to \mathbb{R}$ as above. The Hessian of f is the symmetric 2-tensor:

$$H(f; X, Y) = \nabla_Y \nabla_X f - \nabla_{\nabla_Y X} f.$$

The Laplacian of f is then defined as the trace of the Hessian with respect to the metric:

$$\Delta f = g^{ij} H(f; \partial_i, \partial_j),$$

where g^{ij} stands for the i, j element of the inverse metric matrix and ∂_i is the i-th coordinate vector field.

3. Parametric Estimation

Let $(E, \mathcal{F}, \mu)$ be a measure space. In the sequel, all distributions are assumed to be absolutely continuous with respect to μ and all densities will thus implicitly refer to it. In parametric density estimation, one wants to approximate an unknown distribution with density f by a member of a given parameterized family $\{f_\theta : E \to \mathbb{R}^+, \theta \in \Theta\}$ of densities. Most of the time, the optimal θ^* is found using a maximum likelihood procedure: if $(X_i)_{i=1\ldots N}$ is an iid sample with common distribution f, then

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^{N} f(X_i | \theta).$$

The only requirement on the domain E of the family $\{f_\theta, \theta \in \Theta\}$ is to be a measure space, which of course encompasses the Riemannian manifold case, with $\mu = \text{vol}$, the Riemannian volume.

Being able to obtain a meaningful parameterized distribution family on a general manifold is not an easy task in general. Some clues will be given at the end of this section. In special cases, some well known distributions were introduced. Some of them will be presented now.

3.1. Directional Statistics

Directional statistics [5] deals with inference on unit vectors samples and introduces ad hoc distributions. Since unit vectors can be seen as points on the unit sphere of the underlying vector space, it yields a basic, yet extremely useful example of parameterized families of distributions on a Riemannian manifold.

Since the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ has rotational invariance, it is expected that the parameterized families $(f_\theta)_{\theta \in \Theta}$ exhibit the same behavior, i.e.,

$$\forall A \in \mathrm{SO}(d), \forall \theta \in \Theta, \forall x \in \mathbb{S}^n, f_{A\theta}(Ax) = f_\theta(x). \tag{1}$$

Please note that, to be able to write such a covariance property, it is required to have an action of the group $\mathrm{SO}(d)$ on Θ. The case $\Theta = \mathbb{S}^{d-1}$ is, up to our knowledge, the only one that has been considered by the directional statistics community.

A common choice is the von Mises–Fisher (vMF) distribution on $\mathbb{S}^{d-1}$, denoted $\mathcal{M}(\mu, \kappa)$, given by the density [6]

$$f_\lambda(x; m) = c_d(\kappa)\, e^{\kappa \langle \mu, x \rangle},\ \kappa > 0,\ x \in \mathbb{S}^{d-1}, \tag{2}$$

where

$$c_d(\lambda) = \frac{\lambda^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\lambda)} \tag{3}$$

is a normalization constant with $I_r(\lambda)$ denoting the modified Bessel function of the first kind at order r. The vMF density is unimodal, parameterized by the mean μ and the concentration parameter $\kappa > 0$ that controls the dispersion of the distribution around the mean. The limiting, degenerate case $\kappa = 0$ yields the uniform distribution on the sphere.

When the expectations of the projections on a fixed orthonormal basis $(e_1, \dots, e_d)$ are given, it is a maximum entropy distribution. This fact can be easily seen by writing the associated variational problem with linear constraints:

$$\begin{cases} \operatorname{argmax}_f \int_{\mathbb{S}^{d-1}} f(x) \log f(x) d\sigma(x), \\ \int_{\mathbb{S}^{d-1}} f(x) \langle e_i, x \rangle d\sigma(x) = a_i, \quad i = 1 \dots d, \\ \int_{\mathbb{S}^{d-1}} f(x) d\sigma(x) = 1, \end{cases} \tag{4}$$

with σ the solid angle measure on the sphere. Using Lagrange multipliers $\lambda_1, \dots \lambda_d$ for the first d constraints and c for the last one yields a general form for the solution:

$$f(x; c, \lambda) = \exp\left(c + \langle x, \lambda \rangle\right) \tag{5}$$

with $\lambda = (\lambda_1, \dots, \lambda_d)$. The c is a normalization constant. The remaining ones can be interpreted as mean parameters by normalization, provided $\lambda \neq 0$:

$$\langle x, \lambda \rangle = \|\lambda\| \langle x, \lambda / \|\lambda\| \rangle.$$

The vMF density extends readily to the Stiefel manifold $O(d, p)$ of p-dimensional orthonormal families in $\mathbb{R}^d$ using the same maximum entropy approach, but using projections on the elementary matrices of dimension $d \times p$. The general form of the distribution is then:

$$f(X; M) = c(M) \exp\left(\operatorname{tr} M^t X\right), \; M \in M_{d,p}, \; X \in O(d, p). \tag{6}$$

As in the spherical vMF case, M cannot be interpreted directly as a mean on $O(d, p)$ and some kind of normalization is needed. In order to better understand the behavior of M, it is useful to use its singular values decomposition (SVD) decomposition [7]. Let $M = U \Sigma V^t$ with $U \in M_{d,d}$, $V \in M_{p,p}$, $\Sigma \in M_{d,p}$ and U, V orthogonal matrices. Then:

$$\operatorname{tr} M^t X = \operatorname{tr} V \Sigma^t U X = \operatorname{tr} \Sigma^t U X V. \tag{7}$$

Let C_j, $j = 1 \dots p$, be the columns of the $d \times p$ matrix XV. It comes:

$$\operatorname{tr} M^t X = \sum_{j=1}^{p} \sigma_j \langle U_j, C_j \rangle, \tag{8}$$

where σ_j is the j-th diagonal element of Σ and U_j is the j-th row of U. The net result is thus a product of vMF, after a change of basis given by the matrix V. A rank deficiency in the matrix M indicates a uniform distribution on a subspace, much like in the standard vMF case with $\kappa = 0$. In the matrix case, the limiting case can occur on full subspaces.

A further extension to the Grassmann manifold can be done by quotienting out the density with respect to the $O(p)$ group action [8].

A generalization of maximum entropy distributions with moment constraints to Riemannian manifolds M can be found in [9], where an analogous of the normal law is obtained. The constraints chosen are, in normal coordinates around the mean value:

$$\begin{cases} \int f(x) d\mathrm{vol}(x) = 1, \\ \int x f(x) d\mathrm{vol}(x) = 0, \\ \int x x^t f(x) d\mathrm{vol}(x) = \Sigma, \end{cases} \tag{9}$$

where Σ is a fixed symmetric, positive definite matrix. The resulting density is parameterized by a mean μ and concentration matrix Γ and is expressed as:

$$f_{\mu,\Gamma}(p) \propto \exp\left(-\frac{\log_{\mu}(p)^{t}\Gamma\log_{\mu}(p)}{2}\right), \quad p \in M. \tag{10}$$

As mentioned in [9], the distribution may not be differentiable or even continuous on the cut locus.

Finally, a different construction [10,11] yields a directional distribution on the hyperbolic d-dimensional space. Only the approach of [11] will be detailed here, as it introduces another way to obtain distributions on manifolds using exit points of a Brownian motion with drift. The general idea underlying this approach is to use a submanifold of a well-known model manifold on which the Brownian motion with drift can be constructed. Starting at a fixed origin, a Brownian motion path will intersect the submanifold for the first time at a point, called the exit point. The distribution of the exit points will yield a generalized directional distribution. The original motivation of this construction comes from the exit distribution of a Brownian motion with drift starting at the origin on the unit circle in $\mathbb{R}^2$. The resulting density turns out to be exactly the vMF.

In the hyperbolic space, the Brownian motion is a diffusion with infinitesimal generator:

$$\frac{x_d^2}{2}\left(\sum_{i=1}^{d}\frac{\partial^2}{\partial x_i^2}\right) - \frac{(d-2)x_d}{2}\frac{\partial}{\partial x_d}, \tag{11}$$

where all the coordinates are given in the half-space model of $\mathbb{H}^d$:

$$\mathbb{H}^d = \{x_1,\ldots,x_d \colon x_i \in \mathbb{R},\, i = 1,\ldots,d-1,\, x_d \in \mathbb{R}^+\}.$$

It is convenient to represent the half-space model of $\mathbb{H}^2$ in $\mathbb{C}$, with $z = x + iy$, $y > 0$. The two-dimensional hyperboloid embedded in $\mathbb{R}^3$ associated with $\mathbb{H}^2$ is given by:

$$\{(x_1,x_2,x_3) \colon x_1^2 + x_2^2 - x_3^2 = -1\}. \tag{12}$$

It admits hyperbolic coordinates:

$$x_1 = \sinh(r)\cos(\theta), x_2 = \sinh(r)\sin(\theta), x_3 = \cosh(r), \tag{13}$$

that transforms to the unit disk model as:

$$u = \frac{\sinh(r)\cos(\theta)}{1+\cosh(r)}, v = \frac{\sinh(r)\sin(\theta)}{1+\cosh(r)}, \tag{14}$$

where θ and r are the angular and radius coordinates. Finally, using a complex representation $z = u + iv$ and the Möbius mapping $z \to i(1-z)/(1+z)$, it comes the expression of the half-plane coordinates:

$$x = \frac{\sinh(r)\sin(\theta)}{\cosh(r)+\sinh(r)\cos(\theta)}, y = \frac{1}{\cosh(r)+\sinh(r)\cos(\theta)}. \tag{15}$$

The hyperbolic von Mises distribution is then defined, for a given $r > 0$, as the density of the first exit on the circle of center i and radius r of the hyperbolic Brownian motion starting at i. Its expression is given in [11] (Section 2.2, Propostion 2) as:

$$f_v(r,\theta) = \frac{1}{2\pi P_{-\nu}^{0}(\cosh(r))}\left(\cosh(r)+\sinh(r)\cos(\theta)\right)^{-\nu}, \tag{16}$$

where $P^0_{-\nu}$ is the Legendre function of the first kind with parameters $0, -\nu$, which acts as a normalizing constant to get a true probability density. The parameter ν is similar to the concentration used in the classical von Mises distribution.

3.2. Gaussian-Like Distributions

The maximum entropy distributions introduced above are not the only possible choice for probabilities on manifolds. Another approach may be to mimic the multivariate normal density using the geodesic distance on the manifold. For the space of symmetric positive definite matrices, one can refer to [12], and to [13] for the general case of symmetric spaces.

Let M be a symmetric space [14]. The Gaussian-like density on M with mean μ and variance $\sigma^2 > 0$ is given by:

$$f_{\mu,\sigma}(p) = \frac{1}{Z(\sigma)} \exp\left(-\frac{d^2(\mu, p)}{2\sigma^2}\right), \tag{17}$$

where the normalizing constant Z is independent from μ. This last fact is one of the motivations to use the above definition: the density computation does not require anything more than the geodesic distance evaluation. The basic facts about Gaussian-like distributions on symmetric spaces are given below.

Definition 1. *A Riemannian symmetric space M is a complete Riemannian manifold on which geodesic symmetries exist everywhere and are isometric.*

The above geometric definition fits within a Lie theoretic construction. For the details on such objects, the reader can refer to [15] or for a more complete reference to [16].

Definition 2. *A Riemannian symmetric space M is diffeomorphic to a quotient G/H where G is a Lie connected group and H is a compact Lie subgroup.*

The Lie group view enables the use of integral formulas [17] (pp. 203–209).

Proposition 1. *Let M be a symmetric space of non-compact type, diffeomorphic to G/K and let $\mathfrak{g} = \mathfrak{h} + \mathfrak{a} + \mathfrak{n}$ be the Iwasawa decomposition of the Lie algebra $\mathfrak{g}$. For any $f \in L^1(M)$:*

$$\int_M f(p) d\mathrm{vol}(p) = C \int_H \int_{\mathfrak{a}} f\left(\exp\left(\mathrm{Ad}(h)a\right) \cdot o\right) D(a) da\, dh,$$

with da the Lebesgue measure on $\mathfrak{a}$, dh the normalized Haar measure on H and:

$$D(a) = \prod_{\alpha \in \Sigma^+} \sinh^{m_\alpha}\left(|\alpha(a)|\right),$$

where the product is taken over the set of positive roots Σ^+ and m_α is the dimension of the root space at α. o is a reference point.

Applying the previous formula to the mapping:

$$\tilde{f}\colon p \mapsto \exp\left(-\frac{d^2(\mu, p)}{2}\right)$$

with an origin at μ yields:

$$\int_M \tilde{f}(p) d\mathrm{vol}(p) = C \int_H \int_{\mathfrak{a}} \exp\left(-\frac{B(a, a)}{2\sigma^2}\right) D(a) da\, dh,$$

with B the Killing form. Since the Haar measure dh is normalized and the integrand does not depend on h, it comes:

$$\int_M \tilde{f}(p)d\mathrm{vol}(p) = Z(\sigma) = C\int_{\mathfrak{a}} \exp\left(-\frac{B(a,a)}{2\sigma^2}\right) D(a)da,$$

thus proving that the normalizing constant of the Gaussian-like distribution is independent from μ. Ref. [13] also proposes a maximum-likelihood estimator (MLE) suitable for the estimation of the parameters μ, σ.

Proposition 2. *Let $X_1, \dots, X_n$ be an iid sample drawn from the density $f_{\mu,\sigma}$. The MLE estimator $\hat{\mu}$ (resp. $\hat{\eta}$) of μ (resp. $-1/2\sigma^2$) is the Riemannian barycentre of the sample (resp. $\operatorname{argmax}_\eta \eta\hat{\rho} - \log Z(\sigma(\eta))$). In the expression of the $\hat{\eta}$ estimator, $\hat{\rho}$ is given by:*

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} d^(\hat{\mu}, X_i).$$

3.3. Wrapped Distributions

Among numerous properties, the Gaussian densities in $\mathbb{R}^d$ are known to be solutions of the heat kernel. When the manifold of interest M is obtained as a quotient of a model space H by a discrete group, which is the case for example with Riemann surfaces, the heat kernel on M can be obtained by wrapping the one on H along the orbits of the group action.

The most basic distribution arising that way is the so-called wrapped Gaussian density on the unit circle in $\mathbb{R}^2$. It is defined as:

$$f_{wg}(\theta;\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\sum_{k\in\mathbb{Z}} \exp(-(\theta + 2k\pi)/(2\sigma^2)) \tag{18}$$

and clearly exhibits a period 2π. The parameter σ controls the concentration. It is worth noticing that $f_{wg}(\theta;\sigma)$ is in fact the heat kernel on the circle. Evaluating the density involves finding the sum of a convergent series, which may be costly when the computation is done numerically. In the case of the wrapped Gaussian density, the very fast decay at infinity of the usual normal distribution limits the number of terms to be taken into account.

In [18], the heat kernel of simply connected Riemann surfaces is given by wrapping one of the only three possible model spaces: the Euclidean plane, the hyperbolic plane and the sphere. The respective heat kernels are:

$$K(x,y,t)_{\mathbb{R}^2} = \frac{1}{4\pi t}\exp\left(-\frac{\|x-y\|^2}{4t}\right), \tag{19}$$

$$K(x,y,t)_{\mathbb{H}^2} = \frac{\sqrt{2}e^{-\frac{t}{4}}}{(4\pi t)^{3/2}}\int_{d(x,y)}^{\infty} \frac{se^{-\frac{s^2}{4t}}}{\sqrt{\cosh s - \cosh d(x,y)}}ds, \tag{20}$$

$$K(x,y,t)_{\mathbb{S}^2} = \frac{1}{4\pi}\sum_{n\in\mathbb{N}} (2n+1)\exp(-n(n+1)t)P_n\left(\cos d_{\mathbb{S}^2}(x,y)\right), \tag{21}$$

where the expression for the hyperbolic plane comes from [19] (p. 360). The distances $d(\cdot,\cdot)$ are the geodesic distances.

Theorem 1. *Let M be a Riemann surface, U its universal cover and G its covering group. Let K_U be the heat kernel on U. Then, the heat kernel on M is obtained by wrapping K_U along the orbits of the covering group action:*

$$K_M(x,y,t) = \sum_{g\in G} K_U(\tilde{x}, g\cdot\tilde{y}, t),$$

where $\tilde{x}, \tilde{y}$ are fixed pre-images of respectively x, y for the covering map.

The proof can be found in [18] (pp. 7–8). In principle, Theorem 1 yields a density similar to a directional one, but on a more general class of manifolds. Unfortunately, while a closed-form solution for the kernel K_U is known, and is one of equations (19)–(21), the wrapped kernel is generally only computable numerically, after truncation to a finite number of terms in the sum.

In the case of surfaces with covering space $\mathbb{H}^2$, it is possible to obtain a more convenient description. In this case, the genus g of the surface is strictly larger than 1, and the fundamental region in $\mathbb{H}^2$ is a hyperbolic polygon with $4g$ sides. For any $g \in G$, its length is defined to be $l(g) = \inf_x d(x, gx)$, or using the conjugacy class of g: $l(g) = \inf_x d(x, kgk^{-1})$ where k runs over G. Elements of G with non zero length are conjugate to hyperbolic elements in $SL(2, \mathbb{R})$ and are thus conjugate to a scaling $x \mapsto \lambda^2 x$. Furthermore, a conjugacy class represents a free homotopy class of closed curves, which contains a unique minimal geodesic whose length is $l(g)$, where g is a representative element. For p a primitive element, let G_p denote its centralizer in G. The conjugacy classes in G are all of the form $gp^n g^{-1}, g \in G/G_p$ with p primitive and $n \in \mathbb{Z}$. The wrapped kernel can then be rewritten as:

$$K_M(x, y, t) = \sum_p \sum_{g \in G/G_p} \sum_{n \in \mathbb{Z}} K_{\mathbb{H}^2}(gx, p^n gy, t), \tag{22}$$

where p runs through the primitive elements of G. It indicates that the kernel K_M can be understood as a sum of elementary wrapped kernels associated to primitive elements, namely those $\tilde{k}_p$ defined by:

$$\tilde{k}_p(x, y, t) = \sum_{n \in \mathbb{Z}} K_{\mathbb{H}^2}(x, p^n y, t), \tag{23}$$

with p primitive. Finally, p being hyperbolic, it is conjugate to a scaling, so it is enough to consider kernels of the form:

$$\tilde{k}_p(x, y, t) = \sum_{n \in \mathbb{Z}} K_{\mathbb{H}^2}(x, (\lambda^2)^n y, t), \tag{24}$$

with $\lambda > 1$ a real number. To each primitive element p, a simple closed minimal geodesic loop is associated, which projects onto the axis of the hyperbolic transformation p. In the Poincaré half-plane model, such a loop unwraps onto the segment of the imaginary axis that lies between i and $i\lambda^2$. It is easily seen that the action of the elements $p^n, n \in \mathbb{Z}$ will give rise to a tiling of the positive imaginary axis with segments of the form $[\lambda^{2n}, \lambda^{2(n+1)}[$. This representation allows a simple interpretation of the elementary wrapped kernels $\tilde{k}_p$, where the wrapping is understood as a winding.

Once again, the computational cost involved in the summation may be high. An approximation of the true wrapped kernel is given in [20]. It is similar to the vMF that approximates the wrapped Gaussian density in the circular case.

3.4. Exponential Families Arising from Group Actions

In many physical systems, some quantities must be invariant under the action of a group. This is a consequence of the celebrated Noether theorem [21]. Looking at little bit deeper at this theorem reveals the importance of a mapping, called the momentum map, that turns out to be constant under the evolution of the system. Turning back to densities but keeping the physical framework in mind, it seems natural to find families with a maximum entropy property, with constraints based on the momentum map. This approach and its relationship with information geometry has been thoroughly studied in [22] and will be presented later.

Another view at the same problem is to start from natural exponential families and try to impose group invariance [23]. Let E be a finite dimensional vector space and let E^* be its dual.

Definition 3. *The set $\mathcal{E}$ is the subset of positive Radon measures μ on E such that:*

- *μ is not concentrated on a proper affine subspace of E.*

- *The set of the $\theta \in E^*$ such that:*
$$\int_E \exp\langle \theta, x\rangle \mu(dx) < +\infty$$
has non empty interior, hereafter denoted $\Theta(\mu)$.

Any measure in $\mathcal{E}$ gives rise to a natural exponential family.

Definition 4. *Let $\mu \in \mathcal{E}$. The natural exponential family with base measure μ is the parameterized family $P_\theta, \theta \in \Theta(\mu)$ defined by:*
$$P_\theta(dx) = \exp\left(\langle \theta, x\rangle - C_\mu(\theta)\right)\mu(dx).$$

Given a group G acting on E, a natural exponential family $P_\theta, \theta \in \Theta(\mu)$ with base μ is said to be invariant if for any $g \in G$ and any p_θ, $g.p_\theta = p_{\theta'}$ for a θ' in $\Theta(\mu)$. In the original work, groups of affinities were considered, namely: $g.x = A_g x + v_g, x \in E$, where A_g is the linear part of the affinity and v_g the translation part. The main theorem characterizing the group action invariance for a given natural exponential family is [23]:

Theorem 2. *Let $P_\theta, \theta \in \Theta(\mu)$ be a natural exponential family and G a group of affinities. The family P_θ is invariant under the action of G iff it exist an application $a\colon G \to E^*$ and an application $b\colon G \to \mathbb{R}$ such that:*

$$\forall (g, g') \in G \times G,\ a(gg') = {}^tA_g^{-1} a(g') + a(g), \tag{25}$$
$$\forall (g, g') \in G \times G,\ b(gg') = b(g) + b(g') - \langle a(g'), A_g^{-1} v_g\rangle, \tag{26}$$
$$\forall g \in G\, g \cdot \mu(dx) = \exp\left(\langle a(g), x\rangle + b(g)\right)\mu(dx). \tag{27}$$

When G is a Lie group, a, b are differentiable mappings. With a group theoretic view, a is a cocycle for the action $g\colon x \in E^* \mapsto g \cdot x = {}^tg^{-1} \cdot x$. Theorem 2 was given in [24] and improved in [23] (p. 4). As an example of use, natural exponential families on $\mathbb{R}^d$ are characterized by the next theorem.

Theorem 3. *A natural exponential family on $\mathbb{R}^d$ is invariant under the action of $SO(d)$ iff it admits as base measure $\mu = c\delta_0 + \phi(\nu \otimes \sigma)$, where $c \geq 0$, δ_0 is the delta measure at the origin, σ is the surface measure on $\mathbb{S}^{d-1}$ and:*

- *ν is a measure on $[0, +\infty]$, with $\nu([0,1]) < +\infty$ and it exists $k > 0$ such that:*
$$\int_{[1,+\infty]} x^{-(d-1)/2} \exp(kx)\nu(dx) < +\infty,$$
- *$\phi\colon [0, +\infty] \times \mathbb{S}^{d-1} \to \mathbb{R}^d - \{0\}$ is the polar coordinates mapping: $(r, u) \mapsto ru$.*

Going back to the mechanical formalism and the momentum map, invariant exponential families can be put into a wider framework. The underlying object is a symplectic manifold (M, ω) where ω is a closed, non degenerate two-form on M. Let G be a connected Lie group acting on M.

Definition 5. *The action of G on M is said to be symplectic if for all $g \in G$, $g \cdot \omega = \omega$.*

The group action defines canonical vector fields on TM.

Definition 6. *Let $\xi \in \mathfrak{g}$. The vector field X_ξ is defined by:*

$$X_\xi\colon x \in M \mapsto \left.\frac{d}{dt}\right|_{t=0} \exp_e(t\xi) \cdot x,$$

where e denotes the identity of G.

The vector field X_ξ can be interpreted as the infinitesimal action of g on the points of M.

Definition 7. *A mapping $U\colon M \to \mathfrak{g}^*$ is said to be a momentum map for the G-action if for all $\xi \in \mathfrak{g}$:*

$$d\alpha_\xi = X_\xi \lrcorner \omega,$$

where α_ξ is the 0-form defined as:

$$\forall x \in M,\ \alpha_\xi(x) = \langle U(x), \xi \rangle.$$

As noticed by Souriau [25], the momentum map allows a definition of exponential densities.

Definition 8. *Let (M, ω) be a symplectic manifold of dimension 2d. The 2d-form $\omega^n/n!$ is a volume form on M, called the Liouville form.*

For a symplectic manifold, the Liouville form is the canonical one and will be denoted vol as in the Riemannian case. It is invariant by symplectomorphisms, which is a key ingredient in the definition of exponential families on M.

Definition 9. *Let (M, ω) be a symplectic manifold. Let G be a connected Lie group acting on M with momentum map U. If the set of $\xi \in \mathfrak{g}$ such that:*

$$\int_M \exp\left(-\langle U(x), \xi \rangle\right) d\mathrm{vol}(x) < +\infty$$

has non empty interior, hereafter denoted by Ξ, the exponential family associated with the group action is defined to be:

$$\forall \xi \in \Xi,\ P_\xi(d\mathrm{vol}(x)) = \exp\left(-\langle U(x), \xi \rangle + C(\xi)\right) d\mathrm{vol}(x).$$

The momentum map may be used to define the analog of the usual moments and will in turn allow constraints definition in a maximum entropy approach.

Definition 10. *With the hypothesis of Definition 9, the nth moment of a probability density f on M is defined as:*

$$E_n(f) = \int_M U \otimes^n (x) f(x) d\mathrm{vol}(x).$$

Following [25], the exponential distributions are maximal entropy ones. Assuming that the first and second moments are defined for the exponential family, the next proposition holds.

Proposition 3. *Under the assumptions of Definition 9, the exponential distributions are the one with the largest entropy under the constraint $E_n = K$, with K a fixed vector in $\mathfrak{g}^*$.*

The Souriau approach to invariant Gibbs measure has the obvious advantage of being intrinsic and adapted to a given group action. The parameters of the exponential families are elements of the Lie algebra and must be understood as a general way of fixing a generalized location (please note that it encompasses 'scale' parameters also). It requires a symplectic base manifold, that is very natural in mechanics, but may be a little bit tricky to obtain in a more general setting.

4. Non-Parametric Density Estimation by Projection

4.1. The Euclidean Case

The intuitive idea behind the projection approach to density estimation on measure set $(E, \mathcal{F}, \mu)$ is to use an orthonormal Hilbert basis $(\phi_n)_{n\in\mathbb{N}}$ of the space $L^2(E, \mu)$ to construct an approximation of the unknown density $f \in L^2(E, \mu)$ from its projections. Namely,

$$\alpha_n = \mathbb{E}_f[\phi_n(X)] = \int_E \phi_n(x)f(x)d\mu(x), \tag{28}$$

where $\mathbb{E}_f$ denotes expectation taken with respect to the density f. The reconstruction formula in $L^2(E,\mu)$ then reads as:

$$f : x \mapsto \sum_{n\in\mathbb{N}} \alpha_n \phi_n(x). \tag{29}$$

To turn the expansion into a density estimator, it is necessary to have an estimator of the coefficients α_n. Furthermore, in applications, the series (29) has to be truncated to a finite number of terms. It is thus advisable to have a fast decay of the expansion coefficients when n goes to infinity.

For the first point, an empirical estimator of the expectation is generally used. Assuming an iid sample $(X_i)_{i=1\ldots N}$, the n-th projection is estimated as:

$$\hat{\alpha}_n^N = \frac{1}{N}\sum_{i=1}^{N} \phi_n(X_i), \tag{30}$$

and the density estimator is

$$\hat{f}^N(x) = \sum_{i=1}^{N} \hat{\alpha}_n^N \phi_n(x). \tag{31}$$

Taking the expectation shows that the estimator is unbiased. It is worth to notice that the projection method does not use more than the measure space structure and is thus easy to use on many spaces, including manifolds. It is also very fast to evaluate provided the expansion functions are known and can be computed easily. It nevertheless suffers from two flaws:

- The estimated density $\hat{f}^N$ is not necessarily non-negative as the expansion functions are generally not.
- Depending on the underlying measure space, a countable Hilbert basis may not exist and, even if this holds, the expansion functions may not be expressed in a closed form.

Most of the usual Hilbert spaces are countable. However, the Besocovitch space B^2 of almost periodic functions [26] is a classical example for which an uncountable Hilbert basis exists.

4.2. The Riemannian Case

When the underlying measure space is a compact Riemannian manifold (M,g) of dimension d, equipped with its volume form, the Laplace–Beltrami operator Δ naturally gives rise to a suitable Hilbert basis of $L^2(M,\mathrm{vol})$, which we will denote by $L^2(M)$ for short. Indeed, there exists a sequence $(\lambda_n,\phi_n)_{n\in\mathbb{N}}$ such that for all $n \in \mathbb{N}$,

$$\Delta\phi_n = \lambda_n\phi_n,$$

and the ϕ_n's form a Hilbert basis of $L^2(M)$. The seminal work [27] details the theory of non-parametric projection based estimation in this case. Unfortunately, only in a few cases are the eigenfunctions of Δ known, limiting the applicability of the method. In the sequel, the estimator $\hat{f}^N$ with expansion truncated to the Q lowest terms will be denoted as $\hat{f}_Q^N$. Two main theorems describe the behavior of the estimator.

Theorem 4. *Let f be a density of class $C^s(M)$, with derivatives belonging to $L^2(M)$. For any integer $I > 0$, there exist constants A_f, B_f such that, for all $Q \geq I$,*

$$\mathbb{E}_f\left[\left\|f - \hat{f}_Q^N\right\|_{L^2(M)}^2\right] \leq A_f\frac{Q^{d/2}}{N} + B_f Q^{-s}. \tag{32}$$

The second theorem gives a L^∞ rate:

Theorem 5. *Let f be a density of class $C^s(M)$, $s > d/2$, with derivatives belonging to $L^2(M)$. For any integer $I > 0$, there exist constants A_M, B_f such that, for all $Q \geq I$,*

$$\mathbb{E}_f\left[\left\|f - \hat{f}_Q^N\right\|^2_{L^\infty(M)}\right]^{1/2} \leq A_f \frac{Q^{d/2}}{N^{1/2}} + B_f Q^{(d/2-s)/2}. \tag{33}$$

The proofs are quite technical and the interested reader may refer to [27] for the details.

When the eigenfunctions of the Laplacian are known, the method is quite effective, since the evaluation of the estimated density at a point does not depend on the sample size. In most cases, however, a closed form for the eigenfunctions is not available, thus limiting the practical usability of this estimator.

A possible workaround is to estimate the true eigenfunctions and eigenvalues from an approximate discrete problem. In the approach of [28], a weighted graph is constructed from a net of points on the manifold M, with weights given by a function of the geodesic distance between vertices. For the graph, extracting the eigenfunctions and eigenvalues boils down to a standard linear algebra problem and can thus be solved efficiently. The result of the procedure is a finite set of eigenvectors, which represent discrete measures on the manifold. The projection estimator in such a case yields a quantization (i.e., a discrete approximation) of the estimated measure. Going back to a density can be done using a smoothing procedure, or simply by turning the discrete approximation to a piecewise constant one. In both cases, evaluation at point requires the computation of the geodesic distance to all the samples, thus making the overall procedure far less efficient.

Finally, it is worth mentioning that Laplacian eigenfunctions and representations are closely related when the underlying space is a Lie group. The reader may refer to [29] for the special case $SO(d)$.

5. Non-Parametric Kernel Estimation

Non-parametric estimation of densities is performed using a sum of elementary bell-shaped functions known as kernels most of the time. It was introduced in the 1960s by Parzen in its seminal work [30] and Rosenblatt [31].

5.1. The Euclidean Case

Assuming that the unknown probability density f is univariate, defined on $\mathbb{R}$, it can be estimated using an iid sample $X_i, i = 1 \ldots N$ as:

$$\hat{f}\colon x \in \mathbb{R} \mapsto \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right), \tag{34}$$

where $K\colon \mathbb{R} \to \mathbb{R}^+$ is a symmetric kernel, i.e., a measurable mapping verifying $K(x) = K(-x)$ and integrating to 1,

$$\int_{\mathbb{R}} K(x)dx = 1.$$

The parameter h is a strictly positive real number called the bandwidth of the estimator. It controls the degree of smoothing, and has to be tuned to get the best compromise between smoothness and accuracy. Kernel density estimators are biased. Their bias can be controlled in the following way.

Proposition 4.

$$\left|f(x) - \mathbb{E}[\hat{f}(x)]\right| \leq \int_{\mathbb{R}} K(u)\,|f(x) - f(x - hu)|\,du. \tag{35}$$

Proof. Let X be a random variable with law the common law of the sample. Taking the expectation of $\hat{f}$ gives:

$$\mathbb{E}[\hat{f}(x)] = \sum_{i=1} \frac{1}{Nh}\mathbb{E}\left[K\left(\frac{x-X_i}{h}\right)\right] = \frac{1}{h}\mathbb{E}\left[K\left(\frac{x-X}{h}\right)\right] = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)dy.$$

Letting $u = (x-y)/h$,

$$\frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)dy = \int_{\mathbb{R}} K(u)f(x-hu)du$$

comes. Then, since the kernel integrates to 1,

$$\left|f(x) - \mathbb{E}[\hat{f}(x)]\right| \leq \int_{\mathbb{R}} K(u)\left|f(x) - f(x-uh)\right| du,$$

hence proving the claim. □

When the density is Lipschitz and the kernel satisfies

$$\int_{\mathbb{R}} K(u)|u|du < +\infty,$$

the bound can be improved to:

$$\left|f(x) - \mathbb{E}[\hat{f}(x)]\right| \leq Ch\int_{\mathbb{R}} K(u)|u|du,$$

where C is the Lispchitz constant. As expected, the bias vanishes as h goes to 0 but is non zero when $h > 0$. The variance of the estimator can be computed pretty much the same way. Since the X_i's are independent,

$$\mathbf{var}\left(\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-X_i}{h}\right)\right) = \frac{1}{Nh^2}\mathbf{var}\, K\left(\frac{x-X}{h}\right)$$

comes. Using $f(x-hu) = f(x-hu) - f(x) + f(x)$, we have:

$$\begin{aligned}\frac{1}{Nh^2}\mathbb{E}\left[K\left(\frac{x-X}{h}\right)^2\right] &= \frac{1}{Nh}\int_{\mathbb{R}} K^2(u)f(x-hu)du \\ &= \frac{1}{Nh}\int_{\mathbb{R}} K^2(u)\left(f(x-hu) - f(x)\right)du + \frac{f(x)}{Nh}\int_{\mathbb{R}} K^2(u)du.\end{aligned}$$

Thus, with the above Lipschitz condition,

$$\frac{1}{Nh^2}\mathbb{E}\left[K\left(\frac{x-X}{h}\right)^2\right] \leq \frac{C}{N}\int_{\mathbb{R}} K^2(u)|u|du + \frac{f(x)}{Nh}\int_{\mathbb{R}} K^2(u)du.$$

It then appears that the upper bound of the variance of $\hat{f}$ goes to infinity as h goes to 0 due to the term

$$\frac{f(x)}{Nh}\int_{\mathbb{R}} K^2(u)du.$$

This fact is an expression of the bias-variance dilemma.

When the density f is of class C^2, the bias can be expressed more conveniently as:

$$b(x) = f(x) - \mathbb{E}[\hat{f}(x)] = -\frac{h^2}{2}f^{(2)}(x)\int_{\mathbb{R}} K(u)u^2du,$$

provided

$$V_K = \int_{\mathbb{R}} K(u)u^2 du < +\infty.$$

Please note that V_K is the variance of the kernel, considered as a probability distribution. It is further assumed in the sequel that the kernel is square summable. The following holds:

$$\mathbf{var}\, K\left(\frac{x-X}{h}\right) = h\int_{\mathbb{R}} K^2(u)f(x-uh)du - h^2\left(\int_{\mathbb{R}} K(u)f(x-hu)du\right)^2,$$

so that the variance of the kernel estimator is

$$\mathbf{var}\, \hat{f}(x) = \frac{1}{Nh}\int_{\mathbb{R}} K^2(u)f(x-uh)du - \frac{1}{N}\left(\int_{\mathbb{R}} K(u)f(x-hu)du\right)^2. \tag{36}$$

The mean square error (MSE) of the estimator $\hat{f}$ is the sum of the variance and the squared bias:

$$\mathbb{E}\left[(\hat{f}(x)-f(x))^2\right] = \frac{1}{Nh}\int_{\mathbb{R}} K^2(u)f(x-uh)du - \frac{1}{N}\left(\int_{\mathbb{R}} K(u)f(x-hu)du\right)^2 + \frac{h^4}{4}\left[f^{(2)}(x)V_K\right]^2. \tag{37}$$

The asymptotic MSE is thus:

$$\mathbb{E}\left[(\hat{f}(x)-f(x))^2\right] \underset{N\to\infty, h\to 0}{\sim} \frac{f(x)}{Nh}\int_{\mathbb{R}} K^2(u)du + \frac{h^4}{4}\left[f^{(2)}(x)V_K\right]^2. \tag{38}$$

There exists an optimal value of h, minimizing the previous expression,

$$h_{opt} = \left(\frac{f(x)\|K\|_2^2}{4NA}\right)^{1/5},$$

where

$$A = \left[f^{(2)}(x)V_K\right]^2.$$

This relation is very classical in density estimation and yields a pointwise convergence rate in $o(n^{-4/5})$, slower than the usual $o(n^{-1})$ in the parametric case.

In the multivariate case, two approaches are of common use. In the first one, the multivariate kernel in $\mathbb{R}^d$ is just an d-fold tensor product of univariate kernels:

$$K(x_1, \ldots, x_d) = \prod_{i=1}^{d} K(x_i).$$

The tensor product kernel integrates to 1 by Fubini's theorem; and the density estimator writes as:

$$f(x) = \frac{1}{h^d}\sum_{i=1}^{N} K\left(\frac{x-X_i}{h}\right), \tag{39}$$

where $x \in \mathbb{R}^d$. Apart from the slower convergence rates, things are similar to the univariate case.

Another option is to let the kernel depend on the norm of the difference between x and the X_i. In such a case, starting again with a univariate kernel K, the density estimator is:

$$f(x) = \frac{C}{h^d}\sum_{i=1}^{N} K\left(\frac{\|x-X_i\|}{h}\right), \tag{40}$$

where C is a normalizing constant, such that:

$$C^{-1} = \int_{\mathbb{R}^d} K(\|u\|)du.$$

In this framework, the multivariate kernel pertains to the class of radial basis functions, which has been thoroughly studied in approximation theory [32].

5.2. The Riemannian Case

Extending the previous derivations to manifolds is not direct, since several problems must be addressed. In the case of Riemannian manifolds, the geodesic distance can be used in place of the Euclidean norm, thus making the radial basis kernels more natural than the tensor product ones. A density estimator based on that is due to Pelletier [33]. His approach is briefly described in the sequel.

Let (M, g) be an orientable Riemannian manifold of dimension d, equipped with its volume form vol, and let $K\colon \mathbb{R}_+ \to \mathbb{R}_+$ be a mapping such that:

$$\int_{\mathbb{R}^d_+} K(\|u\|)du = 1, \tag{41a}$$

$$\int_{\mathbb{R}^d_+} u_j K(\|u\|)du = 0,\ j = 1\dots d, \tag{41b}$$

$$\int_{\mathbb{R}^d_+} \|u\|^2 K(\|u\|)du < +\infty, \tag{41c}$$

$$t > 1 \Rightarrow K(t) = 0, \tag{41d}$$

$$\sup_{t\in\mathbb{R}^+} K(t) = K(0). \tag{41e}$$

The main problem is to define what a radial basis function is in the context of manifolds. A similar question was addressed in [34] for the definition of the Laplacian in radial coordinates. The idea is to use the exponential mapping $\exp_p$ at a fixed point $p \in M$ in a ball centered at the origin in T_pM and of radius less than the injectivity radius $r > 0$ at p and to compose K with the distance function to p to get an equivalent of a radial basis function in $\mathbb{R}^d$. However, the volume form is not invariant under the action of the exponential map unless the manifold is flat. In order to keep the integral of the kernel equal to 1, a multiplicative corrective term has to be added.

Definition 11. *Let $p \in M$. For any v in T_pM, let $\gamma_v\colon t \mapsto \exp\left(\frac{tv}{\|v\|}\right)$ and let w_i, $i = 1\dots d$, be Jacobi fields along γ such that $w_i(0) = 0$, for all $i = 1\dots d$, $Dw_1/dt(0) = v/\|v\|$ and $Dw_i/dt(0)$, $i = 1\dots d$, forms an orthonormal basis of T_pM. The volume density function $\theta_p\colon T_pM \to R$ is defined as:*

$$\theta_p(v) = \|v\|^{d-1} \det\left(w_1(\|v\|), \dots, w_d(\|v\|)\right).$$

By abuse of notations, we will also write $\theta_p(x) = \theta_p(\log_p x)$.

The above definition with varying p extends readily to a mapping $\theta\colon TM \to \mathbb{R}$. The exponential map $\exp_p : T_pM \to M$ induces by pullback a volume form $\exp_p^*$ vol on T_pM, and it is worth noticing that θ_p is its density with respect to the Lebesgue measure of the Euclidean structure on T_pM. That is, in normal coordinates,

$$d\exp_p^* \mathrm{vol}(x) = \theta_p(x)dx. \tag{42}$$

Definition 12. *Let K be a kernel function in the sense of equation* (41) *and $R > 0$ be the injectivity radius of M. The radial kernel at $p \in M$ with bandwidth $0 < r < R$ is defined as the mapping $K_{p,r}$:*

$$K_{p,r} \colon x \in M \mapsto \frac{1}{r^d}\frac{1}{\theta_p(x)} K\left(\frac{d(p,x)}{r}\right).$$

Since K has a support in $[0,1]$ by assumption, the above expression will vanish for $x \notin B(p,r)$, where $B(p,r)$ is the geodesic ball of center p and radius r. We will thus never have to bother about a possible vanishing of θ.

Proposition 5. *The radial kernel satisfies:*

$$\int_{B(p,r)} K_{p,r}(x) d\mathrm{vol} = 1.$$

Proof. In normal coordinates at p and from (42), the following comes:

$$\begin{aligned}\int_{B(p,r)=\exp_p B(0,r)} K_{p,r}(x) d\mathrm{vol} &= \int_{B(0,r)} K_{p,r}\left(\frac{\|x\|}{r}\right) \exp_p^* d\mathrm{vol} \\ &= \int_{B(0,r)} \frac{1}{r^d} K\left(\frac{\|x\|}{r}\right) dx_1 \wedge \cdots \wedge dx_d = 1.\end{aligned} \tag{43}$$

□

The next proposition in [33] shows that the kernel $K_{p,r}$ is centered at p in a probabilistic sense.

Proposition 6. *Let $p \in M$ and δ be the supremum of the sectional curvatures of M. Let μ be a probability measure, absolutely continuous with respect to the measure* vol, *and admitting $K_{p,r}$ as density. If $r < inj_p(M)/2$ and, when $\delta > 0$, also $r < \frac{\pi}{4\sqrt{\delta}}$ then p is the unique minimizer of the function*

$$E \colon x \mapsto \int_M d^2(x,p) d\mu(p).$$

The proof can be found in [33] (Propostion 2.2, p. 5) and relies on a computation of the gradient of E and the convexity of a geodesic ball of radius r satisfying the above conditions. Based on the above results, it is natural to choose the following definition for the Riemannian kernel estimator.

Definition 13. *Let X_i, $i = 1 \ldots N$, be an iid sample on the manifold M with common density function f. Let $0 < r < inj(M)$. The kernel density estimator on N points of f with kernel K and bandwidth r is:*

$$\hat{f}_{n,K,r}(x) = \frac{1}{N} \sum_{i=1}^{N} K_{x,r}(X_i). \tag{44}$$

Using Propostion 5, it is easy to see that $\hat{f}_{n,K,r}$ is a probability density on M. The estimator $\hat{f}_{n,K,r}$ is consistent and behaves roughly as in the Euclidean case.

Theorem 6. *Let f be a C^2 probability density on M with bounded first and second derivative. If $0 < r < inj(M)$, then there exists a constant C_f such that:*

$$\mathbb{E}_f\left[\left|\hat{f}_{n,K,r} - f\right|^2_{L^2(M)}\right] \leq C_f\left(\frac{1}{Nr^d} + r^4\right).$$

The proof that will be given here differs slightly from the one given in [33], and relies on the next lemma.

Lemma 1. *Let $\gamma\colon [0,1] \to M$ be a smooth curve of M and let $f\colon M \to \mathbb{R}$ be of class C^{k+1}. Letting $p = \gamma(0), u = \gamma'(0)$, the Taylor expansion at order k of f along γ with integral remainder at p is given by:*

$$f(\gamma(t)) = f(p) + \nabla_u f t + \ldots \nabla_u^k f \frac{t^k}{k!} + \int_0^t \nabla_{\gamma'(y)}^{k+1} f \frac{(1-y)^k}{k!} dy,$$

where ∇ is an affine connection and $\nabla_u f = u(f) = df_p \cdot u$.

Proof. The mapping $\alpha = f \circ \gamma$ is defined from $[0,1]$ to $\mathbb{R}$ and is of class C^{k+1}. It thus admits a Taylor expansion with integral remainder

$$\alpha(t) = f(p) + \sum_{i=1}^{k} \frac{d^i \alpha(0)}{dt^i} \frac{t^i}{i!} + \int_0^t \frac{d^{k+1}\alpha(y)}{dt^{k+1}} \frac{(1-y)^k}{k!} dy.$$

For a smooth function $\phi\colon M \to \mathbb{R}$, $\frac{d}{dt}\phi \circ \gamma(t) = \nabla_{\gamma'(t)}\phi = d\phi(t) \cdot \gamma'(t)$ and the result follows by induction. □

Lemma 2. *Let $\gamma\colon [0,1] \to M$ be a geodesic of M and $f\colon M \to \mathbb{R}$ be of class C^{k+1}. Then, the remainder in Lemma 1 is upper bounded by*

$$C_f \ell(\gamma)^{k+1} \frac{1}{(k+1)!},$$

where C_f is a constant depending on f and $\ell(\gamma)$ is the length of γ.

Proof. Let $\omega \in T^*M$. Then, $\nabla_{\gamma'}\omega \cdot \gamma' = (\nabla_{\gamma'}\omega) \cdot \gamma' + \omega \cdot \nabla_{\gamma'}\gamma' = (\nabla_{\gamma'}\omega) \cdot \gamma'$. Proceeding by induction

$$\nabla_{\gamma'}^{k+1} f = \left(\nabla_{\gamma'(y)}^{k} df \cdot \gamma'\right) = \left(\nabla_{\gamma'(y)}^{k} df\right) \cdot \gamma'.$$

The term $\nabla_{\gamma'(y)}^{k} df$ is k-linear in γ' and involves derivatives of f up to order $k+1$ along γ. Since f is of class C^{k+1} by assumption, it comes

$$\left|\nabla_{\gamma'}^{k+1} f\right| \leq C_f \|\gamma'\|^{k+1}.$$

Since γ is a geodesic, $\|\gamma'\| = d$, where d is the length of γ. The claim follows by integration. □

Proof of Theorem 6. The proof is essentially an adaptation of the Euclidean case. The bias at $x \in M$ is given by

$$b(x) = \int_{B(x,r)} K_{x,r}(y) f(y) d\mathrm{vol}(y) - f(x),$$

and since the kernel integrates to 1,

$$b(x) = \int_{B(x,r)} K_{x,r}(y) f(y) - f(x) d\mathrm{vol}(y).$$

Using normal coordinates at x, it comes

$$b(x) = \int_{B(0,r)} \frac{1}{r^d} K\left(\frac{\|u\|}{r}\right) (f(\exp_x(u)) - f(x))\, du.$$

Using lemma 1, $f(\exp_x(u)) - f(x) = \nabla_u f(0) + R_f(u)$. In coordinates, $u = u^i \partial_i$ and by linearity of the connexion ∇ and symmetry assumption on the kernel

$$\int_{B(0,r)} \frac{1}{r^d} K\left(\frac{\|u\|}{r}\right) \nabla_u f(0) du = \sum_{i=1}^{d} \nabla_{\partial_i} f(0) \int_{B(0,r)} \frac{1}{r^d} K\left(\frac{\|u\|}{r}\right) u_i du = 0.$$

Since the first and second derivatives of f are assumed to be bounded on M, it exists, by Lemma 2, a constant A_f such that $\left|R_f(u)\right| \leq A_f\|u\|^2$, yielding

$$|b(x)| \leq A_f \int_{B(0,r)} \frac{1}{r^d} K\left(\frac{\|u\|}{r}\right) \|u\|^2 = A_f r^2 \int_{B(0,1)} K\left(\|u\|\right) \|u\|^2 du.$$

When the manifold is of finite volume, then the integral of the squared bias over M is bounded by:

$$A_f^2 r^2 \left(\int_{B(0,1)} K\left(\|u\|\right) \|u\|^2 du\right)^2 \operatorname{vol}(M).$$

The case of unbounded volume is still tractable provided the first and second derivatives of f are square-integrable over M. The computation of the variance is very close to what has been done in the Euclidean case. Since the kernel has vanishing first moment, the only term remaining in the variance is the integral of the squared kernel, which equals

$$\operatorname{var}(p) = \int_{B(p,r)} K^2_{p,r}(x) f(x) d\mathrm{vol} = \frac{1}{r^{2d}} \int_{B(p,r)} \frac{1}{\theta_p^2(x)} K^2\left(\frac{d(p,x)}{r}\right) f(x) d\mathrm{vol}. \tag{45}$$

In normal coordinates around p,

$$\int_{B(p,r)} K^2_{p,r}(x) f(x) d\mathrm{vol} = \frac{1}{r^d} \int_{B0,1} \frac{1}{\theta_p(\exp_p(ru))} K^2\left(\|u\|^2\right) f(\exp_p(ru)) du.$$

In contrast with the bias computation, the θ_p does not cancel. If $(p,x) \mapsto \theta_p(p,x)$ is bounded below by a constant $L > 0$ and using the fact that K is bounded above by $K(0)$, the variance is bounded above by

$$\frac{K^2(0)}{Lr^d} \int_{B(0,1)} f(\exp_p(ru)) du.$$

Integrating the variance over M yields

$$\int_M \operatorname{var}(p) dvol \leq \frac{K^2(0)}{Lr^d} \operatorname{vol}(B(0,1)).$$

Summing the variance and the squared bias completes the proof. □

It is worth noticing that, while the final result is essentially the same as in the Euclidean case, assumptions are somewhat stronger: the kernel used must be of compact support and the θ function has to be bounded below. Furthermore, bandwidth has to be small enough to ensure that the exponential mapping is one to one.

5.3. Computing the Kernel in the Riemannian Case

A important feature of the non-parametric kernel estimation in the Euclidean case is the ease of computation. Estimating the density at a given point is done easily using inner products and function evaluations (polynomials in most cases). Furthermore, when the kernel is compactly supported, it is quite simple to avoid summing vanishing terms: a k-d tree [35] structure can be used to allow a quick distance evaluation, thus excluding points that will not contribute to the estimator.

In the Riemannian case, evaluation the kernel $K_{r,p}$ at a point x requires much more computation. First of all, the geodesic distance from x to p must be found, along with the θ_p function. If a shooting algorithm is used for approximating $d(x,p)$, the derivative of the exponential mapping, needed for θ_p, is generally also computed: the overall cost is thus not really different from the geodesic distance evaluation alone. Nevertheless, unless a closed form for the geodesics is known, the computational cost associated with the process is much higher than with Euclidean data.

In some special cases, however, the θ_p function can be obtained in a closed form. It is the case for symmetric spaces, when the roots system of the base Lie group is known: in such a case, the integral formulas in [17] yields directly θ_p.

6. Discrete Density Estimation through Quantization

Arguably the simplest form of estimate, one can choose for an unknown probability measure μ is a discrete density $\hat{\mu}$ with finite support, i.e., a linear combination of Dirac distributions

$$\hat{\mu}(x) = \frac{1}{n}\sum_{i=1}^{n} w_i \delta_{a_i},$$

with weights $w_i, i = 1, \ldots, n$ that sum up to 1. Finding such an approximation is the purpose of quantization. The theory was originally developed for signal compression purposes in the middle of the 20th century, in order to find appropriate ways to discretize a signal. Quantization was founded for probability distributions on Euclidean spaces, but its generalization to Riemannian manifolds presents no particular difficulty (with the necessary assumptions), and so we will present it directly in that more general setting. For further details, we refer the reader to [2] or the survey paper [36].

6.1. Optimal Quantization

Let μ be a probability measure with compact support on a Riemannian manifold (M, g). We assume that M is complete, i.e., that the exponential map at x is defined on the whole tangent space $T_x M$. Then, by the Hopf–Rinow theorem, M is also geodesically complete, that is, any two points $x, y \in M$ can be joined by a geodesic of shortest length and the geodesic distance is everywhere well defined. Optimal quantization addresses the problem of approximating a random variable X with distribution μ by a *quantized* version $q(X)$ where q has an image Γ of cardinal at most n. More precisely, defining

$$\mathcal{Q}_n = \{q : M \to \Gamma \subset M \text{ measurable}, |\Gamma| \le n\},$$

the optimal quantization problem is to find $q \in \mathcal{Q}$ that minimizes the L^p distance between X and $q(X)$, with the following error

$$q^* = \operatorname*{argmin}_{q \in \mathcal{Q}_n} \mathbb{E}_\mu \left[d(X, q(X))^p \right]. \tag{46}$$

A solution to the above minimization problem is called an optimal n-quantizer, and the minimum error is denoted by

$$e_{n,p}(\mu) = \inf_{q \in \mathcal{Q}_n} \mathbb{E}_\mu \left[d(X, q(X))^p \right].$$

Proposition 7. *The search for optimal n-quantizers can be limited to nearest-neighbor projections, i.e.,*

$$e_{n,p}(\mu) = \inf_{\Gamma, |\Gamma| = n} \mathbb{E}_\mu \left[d(X, q_\Gamma(X))^p \right],$$

where $q_\Gamma : M \to \Gamma = \{a_1, \ldots, a_n\}$ is given by

$$q_\Gamma(x) = \sum_{i=1}^{n} a_i \mathbf{1}_{C_i(\Gamma)}(x), \quad x \in M,$$

and $C_i(\Gamma)$ denotes the i^{th} Voronoi cell associated with Γ,

$$C_i(\Gamma) = \{x \in M, d(x, a_i) \le d(x, a_j) \, \forall j \ne i\}.$$

Proof. Any n-quantizer q of image $\Gamma \subset M$ verifies for all $x \in M$, $d(x, q(x)) \ge \inf_{a \in \Gamma} d(x, a)$, with equality if and only if $q(x) = \operatorname{argmin}_{a \in \Gamma} d(x, a)$. Therefore, the optimal quantizer is the projection to the nearest neighbor of Γ. Moreover, if $|\Gamma| < n$ and $|\operatorname{supp}\mu| \geq n$, one easily checks that q can always be improved, in the sense of criteria (46), by adding an element to its image. This means that an optimal n-quantizer has an image of exactly $|\Gamma| = n$ points, and is of the given form. □

The optimal approximation of X is then given by the image $\hat{X} = q_\Gamma(X)$, while the optimal approximation of its distribution μ is given by the pushforward

$$\hat{\mu} = (q_\Gamma)_* \mu = \sum_{i=1}^{n} f[C_i(\Gamma)]\delta_{a_i}, \tag{47}$$

where the atoms $a_1, \ldots, a_n$ are chosen to minimize the *distorsion function*, simply obtained by evaluating the cost function of (46) at $q = q_\Gamma$,

$$F_{n,p}(a_1, \ldots, a_n) = \mathbb{E}_\mu \left\{ \min_{1 \leq i \leq n} d(X, a_i)^p \right\} = \int_M \min_{1 \leq i \leq n} d(x, a_i)^p d\mu(x). \tag{48}$$

Notice that if we seek to approximate μ by a single point $a \in M$ (i.e., $n = 1$) with respect to an L^2 criteria ($p = 2$), we retrieve the definition of the Riemannian center of mass, also called the Fréchet mean [37].

$$\bar{x} = \mathbb{E}_\mu(X) = \operatorname{argmin}_{a \in M} \int_M d(x, a)^2 d\mu(x). \tag{49}$$

It is worth noting that the optimal quantization problem coincides with the optimal transport problem of approximating μ by the closest discrete measure with at most n supporting points, with respect to the L^p–Wasserstein distance.

Proposition 8. *Let $\mathcal{P}_n$ denote the set of all measures ν on M with $|\text{supp}\nu| \leq n$. Then,*

$$e_{n,p}(\mu) = \inf_{\nu \in \mathcal{P}_n} W_p(\mu, \nu)^p,$$

where W_p denotes the Wasserstein distance of order p, i.e.,

$$W_p(\mu, \nu) = \inf_P \int_{M \times M} d(y, z)^p dP(y, z).$$

Here, the infimum is taken over all measures P on $M \times M$ with marginals μ and ν.

The proof in the Euclidean case can be found in [2], and it applies verbatim to measures on manifolds. We restitute it here for the sake of completeness.

Proof. Let $q \in \mathcal{Q}$ and $f : M \times M \to M$, $f(x) = (x, q(x))$ and $g : M \times M \to \mathbb{R}_+$, $g(y, z) = d(y, z)^p$. Then, by definition of the image measure $f_*\mu$, we have

$$\mathbb{E}_\mu \left[d(X, q(X))^p \right] = \mathbb{E}_\mu \left[g \circ f(X) \right] = \mathbb{E}_{(f_*\mu)} \left[g(Y, Z) \right] = \int_M d(y, z)^p d(f_*\mu)(y, z).$$

Noticing that $\int_{y \in M} (f_*\mu)(dy, dz) = \mu(q^{-1}(dz))$ and $\int_{z \in M} (f_*\mu)(dy, dz) = \mu(dy)$, i.e., that $(f_*\mu)$ has marginals q and $q_*\mu$, we get

$$e_{n,p}(\mu) \geq \inf_{q \in \mathcal{Q}_n} W_p(\mu, q_*\mu)^p \geq \inf_{\nu \in \mathcal{P}_n} W_p(\mu, \nu)^p.$$

On the other hand, if $\nu \in P_n$ has support $\Gamma = \{a_1, \dots, a_n\}$, and P has marginals μ and ν, then

$$\begin{aligned}\int_{M\times M} d(y,z)^p P(dy,dz) &= \int_{M\times\Gamma} d(y,z)^p P(dy,dz) \\ &\geq \int_{M\times\Gamma} \min_{1\leq i\leq n} d(y,a_i)^p P(dy,dz) = \int_M \min_{1\leq i\leq n} d(y,a_i)^p \mu(dy),\end{aligned}$$

which gives $W_p(\mu,\nu)^p \geq F_{n,p}(a_1, \dots, a_n)$ and finally

$$\inf_{\nu\in\mathcal{P}_n} W_p(\mu,\nu)^p \geq \inf_{(a_1,\dots,a_n)} F_{n,p}(a_1,\dots,a_n) = e_{n,p}(\mu).$$

□

An important question that arises now is the existence of a minimizer of (48). The proof of the following claim can be found in [38].

Proposition 9. *Let M be a complete Riemannian manifold and μ a probability distribution on M with density and a compact support. Then, the distorsion function $F_{n,p}$ is continuous and admits a minimizer.*

The minimizer $\alpha = (a_1, \dots, a_n)$, referred to as *optimal n-centers*, is in general not unique: any symmetry of μ, if it exists, will transform a minimizer into another minimizer of $F_{n,p}$. For example, any rotation of the optimal n-centers of the uniform distribution on the sphere conserves optimality.

The second question that comes naturally is: how does the error $e_{n,p}(\mu)$ one makes by approximating μ by (47) evolve when the number n of points grows? In the vector case, Zador's theorem [2] (Theorem 6.2) tells us that it decreases to zero as $n^{-p/d}$, and that the limit of $n^{p/d}e_{n,p}(\mu)$ is proportional to the p^{th} *quantization coefficient*, i.e., the limit (which is also an infimum) when μ is the uniform distribution on the unit square of $\mathbb{R}^d$

$$C_p\left([0,1]^d\right) = \lim_{n\geq 1} n^{p/d} e_{n,p}\left\{\mathcal{U}\left([0,1]^d\right)\right\}.$$

Moreover, when μ is absolutely continuous with density h, the asymptotic empirical distribution of the optimal n-centers is proportional to $h^{d/(d+p)}$.

In the case of a Riemannian manifold M, the moment condition of the flat case generalizes to a condition involving the curvature of M. The following term measures the maximal variation of the exponential map at $x \in M$ when restricted to a $(d-1)$-dimensional sphere $S_\rho \subset T_xM$ of radius ρ

$$A_x(\rho) = \sup_{v\in S_\rho, w\in T_vS_\rho, \|w\|=\rho} \left\|d_v \exp_x(w)\right\|.$$

The following generalization of Zador's theorem to Riemannian quantization was proposed by Iacobelli [39] (Theorem 1.4 and Corollary 1.5) .

Theorem 7. *Let M be a complete Riemannian manifold without boundary, and let $\mu = h\,\mathrm{dvol} + \mu_s$ be a probability measure on M, where dvol denotes the Riemannian volume form and μ_s the singular part of μ. Assume there exist $x_0 \in M$ and $\delta > 0$ such that*

$$\int_M d(x,x_0)^{p+\delta} d\mu(x) + \int_M A_{x_0}\{d(x,x_0)^p\} d\mu(x) < \infty.$$

Then,

$$\lim_{n\to\infty} n^{p/d} e_{n,p}(\mu) = C_p\left([0,1]^d\right) \|h\|_{d/(d+p)},$$

where $\|\cdot\|_r$ denotes the L^r-norm. In addition, if $\mu_s = 0$ and $(a_1, \ldots, a_n)$ are optimal n-centers, then

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{a_i} \xrightarrow{D} \lambda h^{d/(d+p)}\mathrm{d}x \quad \text{as } n \to \infty,$$

where $\xrightarrow{D}$ denotes convergence in distribution and λ is the appropriate normalizing constant.

6.2. A Numerical Scheme

In practice, to compute the optimal n-centers $\alpha = (a_1, \ldots, a_n)$ from potentially large, manifold-valued datasets, one can search for the critical points of the distorsion function. Assume that the only knowledge that we have of the probability measure μ that we want to approximate is through an online sequence of i.i.d. observations $X_1, X_2, \ldots$ sampled from μ. A classical algorithm used for quadratic ($p = 2$) vector quantization, and easily generalized to the Riemannian setting, is the *Competitive Learning Vector Quantization* algorithm, a stochastic gradient descent method based on the differentiability of the distorsion function $F_{n,2}$.

Proposition 10. *Let $\alpha = (a_1, \ldots, a_n) \in M^n$ be an n-tuple of pairwise distinct components and $p > 1$. Then, $F_{n,p}$ is differentiable and its gradient in α is*

$$\operatorname{grad}_\alpha F_{n,p} = \left(-p\int_{\mathring{C}_i(\alpha)} \|\overrightarrow{a_i x}\|^{p-1}\frac{\overrightarrow{a_i x}}{\|\overrightarrow{a_i x}\|}\,\mu(\mathrm{d}x)\right)_{1\le i\le n} \in T_\alpha M^n,$$

where $\mathring{C}_i(\alpha)$ is the interior of the i^{th} Voronoi cell of α and $\overrightarrow{xy} := \exp_x^{-1}(y)$ denotes the vector that sends x on y through the exponential map. In particular, the gradient of the quadratic distorsion function is given by

$$\operatorname{grad}_\alpha F_{n,2} = \left(-2\int_{\mathring{C}_i} \overrightarrow{a_i x}\,\mu(\mathrm{d}x)\right)_{1\le i\le n} = -2\left(\mathbb{E}_\mu \mathbf{1}_{\{X\in\mathring{C}_i\}}\overrightarrow{a_i X}\right)_{1\le i\le n}. \tag{50}$$

The proof can be found in [38].

Notice that optimal n-centers are Riemannian centers of mass of their Voronoi cells, in the sense of (49). More generally, for any value of p, each a_i, $i = 1, \ldots, n$, is the p-mean of its Voronoi cell, i.e., the minimizer of

$$a \mapsto \int_{\mathring{C}_i(\alpha)} d(x,a)^p \mu(\mathrm{d}x).$$

Therefore, the optimal n-centers are always contained in the compact support of μ. Notice also that the opposite direction of the gradient is, on average, given by the vectors inside the expectation. Competitive learning quantization consists in following this direction at each step k, that is, updating only the center a_i corresponding to the Voronoi cell of the new observation X_k, in the direction of that new observation. In the Riemannian setting, instead of moving along straight lines, we simply follow geodesics using the exponential map. This gives a convergent algorithm, as shown in [38], which is particularly adapted to large datasets as it is online, i.e., it processes one data point at a time. Moreover, unlike kernel based methods, it requires few distance computations: of order $n \times N$ instead of N^2 if N is the size of the dataset and $n \ll N$ the size of the summary.

Finding the right number of centers may be a difficult question when there is no a priori knowledge on the distribution to be approximated. In practice, it is mainly a trial and error procedure, unless the problem itself allows an initial guess of the value. An alternative approach is to use the fact that any quantization defines a clustering by Voronoï cells: the quality of the later can be used to assess the performance of the former. Quite a lot of standard indicators exist for that purpose [40]. Just to mention a simple one, the Silhouette [41] is easy to compute and does not require the knowledge of the membership labels.

7. Some Open Problems

Density estimation on manifolds is still an open area of research and as such has several questions that are not yet been answered. Some of them are given below, which may serve as a starting basis for further research.

7.1. Parametric Estimation and Symplectic Structure

A non-trivial aspect of the parametric estimation on manifold is that the parameter space is generally different from the base manifold. In fact, it is already the case for location-scale models in $\mathbb{R}^d$, since the scale parameter is an element of $\mathbb{R}^+$, but the underlying abelian group structure makes the location part an element of $\mathbb{R}^d$, which is also the base space. If one tries to extend the concept of location parameters to manifolds, two approaches can be used:

- Use local coordinates as parameters, and mimic the vector case as in [9].
- Replace the abelian group underlying $\mathbb{R}^d$ by a Lie group acting on the base manifold.

In view of the general results mentioned in Section 3.4 for exponential families defined on symplectic manifolds, the second setting is the most natural one, but is not as general as the first one. However, given a Riemannian manifold M, its cotangent bundle TM^* has a canonical symplectic structure, obtained from the so-called tautological one form [42] (pp. 9–14).

Definition 14. *Let $\pi\colon TM^* \to M$ be the canonical projection and $d\pi\colon TTM^* \to TM$ its derivative. The tautological one-form $\alpha\colon TTM^* \to \mathbb{R}$ is defined pointwise as:*

$$\alpha_\omega(v) = \omega(d\pi v),$$

where $v \in TTM^*$ and $\tilde{\pi} v = \omega$, with $\tilde{\pi}\colon TTM^* \to TM^*$ the canonical projection.

The tautological one-form gives rise to a symplectic form:

Proposition 11. *The two-form $\omega = -d\alpha$ provides TM^* with a symplectic struture. In local coordinates $(x_, \dots x_d, \xi_1, \dots, \xi_d)$, it reads as:*

$$\omega = \sum_{i=1}^{d} dx_i \wedge d\xi_i.$$

Given the symplectic structure on TM^*, one can derive exponential families provided a group action by symplectomorphisms exists. This construction may be a starting point for a quite general definition of parameterized families on manifolds, by moving from the base manifold to its cotangent bundle. It is worth noticing that location models may also be constructed using generating functions [42].

7.2. Manifolds with Boundaries

All the density estimators presented before where defined on manifolds without boundaries. However, in several settings, boundaries arise naturally as limiting cases: as an example, the space of symmetric positive semi-definite matrices of dimension $d \times d$ has stratified boundaries, namely the manifolds of matrices of rank $0 \leq p < d$. Densities localized on the boundaries are degenerate versions of the one defined in the interior, but must be taken into account as they carry a non-vanishing mass. Apart from the projection estimator that fits in the manifolds with boundaries framework, all the other methods must be adapted. In particular, distributions defined on matrix spaces must be parameterized in such a way that rank deficiency is allowed in the parameter space. For manifolds with corners [43], this is easily obtained from the particular structure of the local charts that map open subsets of the manifold to open subsets of $[0, +\infty[\times \mathbb{R}^{d-k}, 0 \leq k \leq d$ and it turns out that all examples of matrix manifolds fall within this frame.

Extending parameterized density estimators or kernel estimators to manifolds with corners will have many practical applications, especially when dealing with matrix statistics. The same applies to optimal quantization, where dirac densities located on the boundaries must be added to the initial model.

7.3. Constrained Quantization

In the Riemannian quantization problem, an approximate distribution in the form $\sum_{i=1}^{n} \mu_i \delta_{a_i}$, with $a_1, \ldots a_n$ the optimal centers located on the base manifold M and $\mu_1, \ldots, \mu_n$ be positive real numbers summing to 1. While such a representation is very natural both from a optimal approximation and clustering point of view, some specific applications require putting additional constraints on the weights $\mu_i, i = 1 \ldots n$ [44]. A common one is to impose that they are all equal, which in a clustering application means that all classes will have an equal expected number of members. The stochastic gradient algorithms presented above is no longer valid for dealing with this problem and has to be adapted. It is still an open question to find a suitable procedure.

8. Conclusions

Probability densities approximation and estimation on Riemannian manifolds are topics receiving an increasing interest in the statistics community. In many applications, data are living on non-Euclidean spaces and adapted procedures must be designed.

In Euclidean spaces, parametric and non-parametric estimation procedures have been intensively studied and are well understood. In the Riemannian manifold setting, several non-equivalent extensions are possible, making the task to find the right one quite tricky. The most obvious way of dealing with manifold valued data is to use a local linearization, like the use of normal coordinates. When applied to densities, it may be used to derive maximum entropy distributions with fixed moments, but some care must be taken with the cut locus, which may be charged by the defined distribution. Other approaches rely on a direct use of the geodesic distance, both in the parametric and non-parametric estimation framework. The computational cost associated with this operation may be high, as a differential problem with boundary conditions has to be solved. When dealing with kernel estimation, Jacobi fields must also be evaluated. While it involves only a classical Ordinary differential equation (ODE). integration, it is an increase in the overall complexity of the procedure.

Parametric estimation using exponential families based on group action invariance are theoretically appealing and makes use of the underlying information on the data. This a especially important when data is highly structured. A closed form momentum map is nevertheless a prerequisite to derive an efficient implementation.

Finally, projection based estimation is in principle very efficient, provided one can obtain the eigenfunctions of the Laplace–Beltrami operator in a closed form, or at least efficiently approximate them. In low dimension, numerical schemes may be used for that purpose, but do not scale well.

Classical directional densities and their generalizations are numerically appealing and offer a sound framework for some manifolds. They are designed on an ad hoc basis, and may not be adapted to all cases. Most of them are maximal entropy based and often exhibit a group invariance. An interesting question is whether approximate directional densities can be found for a wrapped distribution arising from a model heat kernel.

As a general conclusion, extending the usual distributions to general manifolds is by no way an elementary procedure. Furthermore, where in the Euclidean case some distributions satisfy many equivalent defining properties, it will not be the case for manifolds. Maximum entropy is generally a good criterion, provided the fixed moments are defined in a simple and natural way. Finally, an often overlooked issue with densities defined on Riemannian manifolds is the associated computational cost when closed forms are unknown. Since all distance computations require solving a boundary value problem, the complexity of the manifold algorithms may be some order of magnitudes higher than their Euclidean counterparts. This also limits the practical dimension of the problems.

Author Contributions: A.l.B. has contributed to the Riemannian geometry section and the quantization. S.P. has contributed the survey on parametric and non-parametric estimation.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

1. DasGupta, A. *Asymptotic Theory of Statistics and Probability*; Springer Texts in Statistics; Springer: New York, NY, USA, 2008.
2. Graf, S.; Luschgy, H. *Foundations of Quantization for Probability Distributions*; Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 2007.
3. Chern, S.; Smith, F.; de Rham, G. *Differentiable Manifolds: Forms, Currents, Harmonic Forms*; Grundlehren der Mathematischen Wissenschaften; Springer: Berlin/Heidelberg, Germany, 2012.
4. Willmore, T. *Riemannian Geometry*; Oxford Science Publications, Clarendon Press: Oxfordshire, UK, 1996.
5. Mardia, K.V. Statistics of Directional Data. *J. R. Stat. Soc. Ser. B (Methodol.)* **1975**, *37*, 349–393. [CrossRef]
6. Mardia, K.; Jupp, P. *Directional Statistics*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2009.
7. Golub, G.; Van Loan, C. *Matrix Computations*; Johns Hopkins Studies in the Mathematical Sciences; Johns Hopkins University Press: Baltimore, MD, USA, 1996.
8. Chikuse, Y. *Statistics on Special Manifolds*; Lecture Notes in Statistics; Springer: Berlin/Heidelberg, Germany, 2003.
9. Pennec, X. Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *J. Math. Imaging Vis.* **2006**, *25*, 127–154. [CrossRef]
10. Barndorff-Nielsen, O. Hyperbolic Distributions and Distributions on Hyperbolae. *Scand. J. Stat.* **1978**, *5*, 151–157.
11. Gruet, J.C. A Note on Hyperbolic von Mises Distributions. *Bernoulli* **2000**, *6*, 1007–1020. [CrossRef]
12. Said, S.; Bombrun, L.; Berthoumieu, Y.; Manton, J.H. Riemannian Gaussian Distributions on the Space of Symmetric Positive Definite Matrices. *IEEE Trans. Inf. Theory* **2017**, *63*, 2153–2170. [CrossRef]
13. Said, S.; Hajri, H.; Bombrun, L.; Vemuri, B.C. Gaussian Distributions on Riemannian Symmetric Spaces: Statistical Learning with Structured Covariance Matrices. *IEEE Trans. Inf. Theory* **2018**, *64*, 752–772. [CrossRef]
14. Terras, A. *Harmonic Analysis on Symmetric Spaces and Applications I*; Springer: New York, NY, USA, 2012.
15. Duistermaat, J.; Kolk, J. *Lie Groups*; Universitext; Springer: Berlin/Heidelberg, Germany, 1999.
16. Knapp, A.W. *Lie Groups Beyond an Introduction*; Progress in Mathematics; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
17. Helgason, S. *Groups and Geometric Analysis: Integral Geometry, Invariant Differential Operators, and Spherical Functions*; Mathematical Surveys And Monographs; American Mathematical Society: Providence, RI, USA, 2000.
18. Jones, T.H.; Kucerovsky, D. Heat Kernel for Simply-Connected Riemann Surfaces. *arXiv* **2010**, arXiv:1007.5467.
19. McKean, H.P. An upper bound to the spectrum of Δ on a manifold of negative curvature. *J. Differ. Geom.* **1970**, *4*, 359–366. [CrossRef]
20. Nicol, F.; Puechmorel, S. Von Mises-Like Probability Density Functions on Surfaces. In *Geometric Science of Information*; Nielsen, F., Barbaresco, F., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 701–708.
21. Noether, E. Invariant variation problems. *Transp. Theory Stat. Phys.* **1971**, *1*, 186–207. [CrossRef]
22. Barbaresco, F. Koszul Information Geometry and Souriau Geometric Temperature/Capacity of Lie Group Thermodynamics. *Entropy* **2014**, *16*, 4521–4565. [CrossRef]
23. Casalis, M. Familles Exponentielles Naturelles sur R^{d} Invariantes par un Groupe. *Int. Stat. Rev.* **1991**, *59*, 241–262. [CrossRef]
24. Barndorff-Nielsen, O.; Blæsild, P.; Jensen, J.L.; Jørgensen, B. Exponential Transformation Models. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **1982**, *379*, 41–65. [CrossRef]

25. Souriau, J.; Cushman, R.; Vries, C.; Tuynman, G. *Structure of Dynamical Systems: A Symplectic View of Physics*; Progress in Mathematics; Springer Science + Business Media: Berlin/Heidelberg, Germany, 1997.
26. Besicovitch, A. *Almost Periodic Functions*; Dover Edition; Dover Publications: Dover, DE, USA, 1954.
27. Hendriks, H. Nonparametric Estimation of a Probability Density on a Riemannian Manifold Using Fourier Expansions. *Ann. Stat.* **1990**, *18*, 832–849. [CrossRef]
28. Burago, D.; Ivanov, S.; Kurylev, Y. A graph discretization of the Laplace–Beltrami operator. *J. Spectr. Theory* **2014**, *4*, 675–714. [CrossRef]
29. Kim, P.T. Deconvolution density estimation on SO(N). *Ann. Stat.* **1998**, *26*, 1083–1102. [CrossRef]
30. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076. [CrossRef]
31. Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.* **1956**, *27*, 832–837. [CrossRef]
32. Buhmann, M. *Radial Basis Functions: Theory and Implementations*; Cambridge Monographs on Applied and Computational Mathematics; Cambridge University Press: Cambridge, UK, 2003.
33. Pelletier, B. Kernel density estimation on Riemannian manifolds. *Stat. Probab. Lett.* **2005**, *73*, 297–304. [CrossRef]
34. Berger, M.; Gauduchon, P.; Mazet, E. *Le Spectre d'une Variete Riemannienne*; Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 1971.
35. Bentley, J.L. Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM* **1975**, *18*, 509–517. [CrossRef]
36. Pagès, G. Introduction to vector quantization and its applications for numerics. *ESAIM Proc. Surv.* **2015**, *48*, 29–79. [CrossRef]
37. Fréchet, M. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré* **1948**, *10*, 215–310. (In French)
38. Le Brigant, A.; Puechmorel, S. Optimal Riemannian quantization with an application to air traffic analysis. *arXiv* **2018**, arXiv:1806.07605.
39. Iacobelli, M. Asymptotic quantization for probability measures on Riemannian manifolds. *ESAIM Control Optim. Calculus Var.* **2016**, *22*, 770–785. [CrossRef]
40. Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J.M.; Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit.* **2013**, *46*, 243–256. [CrossRef]
41. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
42. Da Silva, A. *Lectures on Symplectic Geometry*; Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 2004.
43. Joyce, D. *On Manifolds with Corners*; Advances in Geometric Analysis, International Press: Boston, MA, USA, 2012; Volume 21, pp. 225–258.
44. Kämpke, T. Constrained quantization. *Signal Process.* **2003**, *83*, 1839–1858. [CrossRef]