Wrapped Gaussian Process Regression on Riemannian Manifolds

Anton Mallasto Aasa Feragen
Department of Computer Science
University of Copenhagen
{mallasto,aasa}@di.ku.dk

Abstract

Gaussian process (GP) regression is a powerful tool in non-parametric regression providing uncertainty estimates. However, it is limited to data in vector spaces. In fields such as shape analysis and diffusion tensor imaging, the data often lies on a manifold, making GP regression non-viable, as the resulting predictive distribution does not live in the correct geometric space. We tackle the problem by defining wrapped Gaussian processes (WGPs) on Riemannian manifolds, using the probabilistic setting to generalize GP regression to the context of manifold-valued targets. The method is validated empirically on diffusion weighted imaging (DWI) data, directional data on the sphere and in the Kendall shape space, endorsing WGP regression as an efficient and flexible tool for manifold-valued regression.

1. Introduction

Regressing functions from Euclidean training data $\{(x_i, y_i)\}_{i=1}^N$ is well studied. Manifold-valued y_i, on the other hand, pose difficulties due to the lack of the vector space structure: Euclidean statistics do not respect the intrinsic structure of manifold-valued data, and the product of inference might not belong to the object category of the data. For example, see Fig. 1, where Gaussian process regression escapes the 2-sphere.

Sometimes the data observed is uncertain. In this case, it is favorable to estimate a distribution over possible regressed functions, yielding uncertainty estimates of the resulting inference. Gaussian process (GP) regression achieves this in a tractable manner. Furthermore, GP regression is an example of Bayesian inference, where it is possible to incorporate prior knowledge to aid the inference. These qualitative properties motivate us to generalize GP regression to Riemannian manifolds.

Related work. Fletcher [1] generalized linear regression to handle manifold-valued data with real covariates by *geodesic regression*; this was later extended to include multi-dimensional covariates [2]. Prior work also consider uncertainty estimates for geodesic regression; by a Kalman filter approach [8] and by stochastic development [7].

Manifold-valued data, however, does not always follow a geodesic trend. Approaches for this non-geodesic setting include kernel-based approaches [5, 6] and a generalization of polynomial regression [4]. Unfortunately, these models do not provide uncertainty estimates.

Improving on this, Pigoli et al. [3] consider a kriging (GP regression) method. The method uses multivariate geodesic regression to form a reference coordinate system, which is used to compute residuals of the manifold-valued data points. Regular GP regression is then applied on the resid-

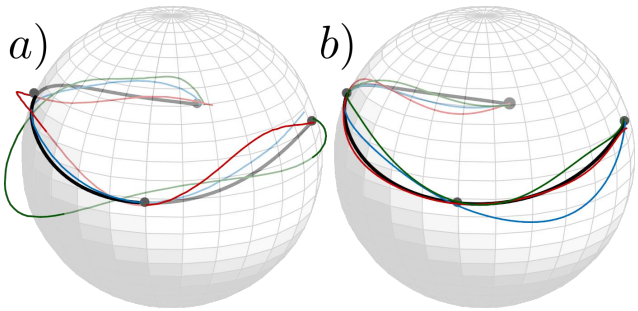

Figure 1. **Why geometrically intrinsic regression is important.** Consider data points (black) on the 2-sphere. In a), we apply ordinary GP regression. The black curve is the prediction and the colorful curves are samples from the predictive distribution, which clearly escape the sphere. In b), we visualize the result using WGP regression, which respects the geometrical constraints of the data.

Method	Non-geod.	Priors	Uncert.	Global
Geod. reg.[1, 2, 3]	No	No	No	Yes
Poly. reg.[4]	Yes	No	No	Yes
Mani. Kriging [3]	Yes	Yes	Yes	No
Kernel reg. [5, 6]	Yes	Yes	No	Yes
Stoch. dev. [7]	No	No	Yes	Yes
Hong et al.[8]	No	Yes	Yes	Yes
WGP reg.	Yes	Yes	Yes	Yes

Table 1. Qualitative comparison of manifold regression models mentioned in this paper. *Global* means, that regression is not carried out in a single tangent space, uncert. is short for uncertainty and Non-geod short for non-geodesic.

uals and the result is mapped back onto the manifold. The procedure, however, depends heavily on the localization of the problem to a single tangent space, and does not offer an intrinsic probabilistic interpretation. Relying on WGPs, our method offers interpretability, and the prior basepoint function used in WGP regression allows avoiding being too local. Furthermore, the kriging method in [3] took advantage of the geodesic submanifold regression to initialize a reference coordinate system. Our method, enables one to take advantage of more general priors, including the use of geodesic submanifold regression.

Steinke and Hein [9] consider the problem of approximating a function between manifolds via minimizing regularized empirical risk. In this setting, also the independent variables are manifold-valued. The WGP regression proposed in this paper can be extended to this setting, as long as a kernel can be defined on the domain, carrying on all the advantages of WGPs mentioned.

Wrapped Gaussian processes appear in directional statistics [10], where a wrapped normal distribution is defined on a 1-sphere S^1, which is then generalized to a multivariate version, and this is then used to define a WGP. This is a special case of our setting, when the manifold is chosen to be the torus $S^1 \times S^1 \times ... \times S^1$.

The contribution can be summarized as follows: We generalize GPs to Riemannian manifolds as wrapped Gaussian procesess (WGPs), and provide a novel framework for non-parametric regression with uncertainty estimates using WGP regression. We demonstrate the method in Section 5 on the 2-sphere by considering a toy example and orientations of the left femur of a walking person, on the manifold of symmetric positive definite matrices for DTI upsampling, and on Kendall shape space, using a data set of Corpus Callosum shapes. The method is analytically tractable for manifolds with infinite injectivity radius, such as manifolds with non-positive curvature. Otherwise, we suggest the approximation in Remark 2. Computationally, the method is relatively cheap, as the only addition compared to GP regression is a single application of the logarithmic map per data point and single exponential map per predicted point.

2. Preliminaries

We briefly summarize the mathematical prerequisities needed. First, we recall how GPs are used in non-parametric regression in the Euclidean case, after which we turn to basic concepts in Riemannian geometry and briefly discuss geodesic submanifold regression.

2.1. Gaussian process regression

Denote by $\mathcal{N}(\mu, \Sigma)$ the multivariate Gaussian distribution with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$, and write the probability density function p as $p(v) = \mathcal{N}(v|\mu, \Sigma)$ for $v \in \mathbb{R}^n$.

A *Gaussian process* (GP) [11] is a collection f of random variables, such that any finite subcollection $(f(\omega_i))_{i=1}^N$ has a joint Gaussian distribution, where $\omega_i \in \Omega \subset \mathbb{R}^l$, and Ω is the *index set*. A GP is entirely characterized by the pair

$$m(\omega) = \mathbb{E}\left[f(\omega)\right], \tag{1}$$

$$k(\omega, \omega') = \mathbb{E}\left[(f(\omega) - m(\omega))(f(\omega') - m(\omega'))^T\right], \tag{2}$$

where m and k are called the *mean function* and *covariance function*, respectively. We denote such a GP by $f \sim \mathcal{GP}(m, k)$. It follows from the definition that the covariance function (*kernel*) k is symmetric and positive semidefinite.

Let $\boldsymbol{D} = \{(x_i, y_i) \mid x_i \in \boldsymbol{x} \subset \mathbb{R}^l,\ y_i \in \boldsymbol{y} \subset \mathbb{R}^n\}$ be the training data. The GP predictive distribution for outputs $\boldsymbol{y}_*$ at the test inputs $\boldsymbol{x}_*$, given in vector form, is

$$p(\boldsymbol{y}_*|\boldsymbol{D}, \boldsymbol{x}_*) = \mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*), \tag{3}$$

$$\boldsymbol{\mu}_* = \boldsymbol{k}_*^T(\boldsymbol{k} + K_{\text{err}})^{-1}\boldsymbol{y}, \tag{4}$$

$$\Sigma_* = \boldsymbol{k}_{**} - \boldsymbol{k}_*^T(\boldsymbol{k} + K_{\text{err}})^{-1}\boldsymbol{k}_*, \tag{5}$$

where, given a kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ we use the notation $\boldsymbol{k} = k(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{k}_* = k(\boldsymbol{x}, \boldsymbol{x}_*)$, $\boldsymbol{k}_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$ and K_{err} is the measurement error variance. In the notation above, the function and k is applied elementwise on the vectors $\boldsymbol{x}, \boldsymbol{x}_*$.

Typically in model selection, the kernel k is picked from a parametric family $\{k_\theta|\theta \in \Theta\}$ of covariance functions, such as the *radial basis function* (RBF) kernels

$$k_{\sigma^2,\lambda}(x, y) = \sigma^2 \exp\left(-\frac{\|x - y\|^2}{2\lambda}\right),\ \sigma^2, \lambda > 0, \tag{6}$$

choosing the parameters (σ^2, λ) so that the *marginal likelihood* $\mathbb{P}\{\boldsymbol{y}|(\sigma^2, \lambda)\}$ is maximized.

2.2. Riemannian geometry

To fix notation, we briefly present the essentials of Riemannian geometry. For a thorough presentation, see [12]. A *Riemannian manifold* is a smooth manifold M with a smoothly varying inner product $g_p(\cdot, \cdot)$ (we will often use the notation $\langle \cdot, \cdot \rangle_p$) on the tangent space T_pM at each $p \in M$, called a *Riemannian metric*, inducing the distance function d between points on the M. Each element (p, v) in the tangent bundle $TM = \bigcup_{p\in M}(p \times T_pM)$ defines a geodesic γ (a curve locally minimizing distance between two points) on M, so that $\gamma(0) = p$ and $\frac{d}{dt}\gamma(t) \mid_{t=0} = v$. The *exponential map* $\text{Exp} : TM \to M$ given by $(p, v) \mapsto \text{Exp}_p(v) = \gamma(1)$, where γ is the geodesic corresponding to (p, v). The exponential map Exp_p at p is a diffeomorphism between a neighborhood $0 \in U \subset T_pM$ and neighbourhood $p \in V \subset M$, which is chosen in a maximal way, so if $V \subsetneq V'$, then a diffeomorphism between V' and a neighborhood in the tangent space cannot be defined anymore. We also call V the *area of injectivity*.

We can define the inverse map $\mathrm{Log}_p : V \to T_pM$, characterized by $\mathrm{Exp}_p(\mathrm{Log}_p(p')) = p'$. Outside of V, we use $\mathrm{Log}_p(p')$ to denote a smallest $v \in T_pM$ chosen in a *measurable*, consistent way. We call the the minimun distance from p to the boundary of a maximal V the *injectivity radius* of Exp_p and the complement of V in M the *cut-locus* at p denoted by $\mathcal{C}_p$. The manifolds with non-positive curvature form an important class of manifolds with infinite injectivity radius, that is, they have an empty cut-locus $\mathcal{C}_p$ for every $p \in M$.

Let M_i be Riemannian manifolds with metrics g_i, exponential maps Exp^i and logarithmic maps Log^i for $i = 1, 2$. Then $M = M_1 \times M_2$ turns into a Riemannian manifold when endowed with the metric $g = g_1 + g_2$, which has the component-wise computed exponential map $\mathrm{Exp}_{(p_1,p_2)}((v_1, v_2)) = \left(\mathrm{Exp}^1_{p_1}(v_1), \mathrm{Exp}^2_{p_2}(v_2)\right)$, akin to the logarithmic map Log on the product manifold.

2.2.1 Probabilistic notions

Let X be a random point on a Riemannian manifold M, the set

$$\mathbb{E}[X] := \left\{ p \mid p \in \arg\min_{q \in M} (\mathbb{E}[d(q, X)^2]) \right\}. \tag{7}$$

is called the *Fréchet means* of X. If there is a unique mean $\bar{p}$, then by abuse of notation we write $\mathbb{E}[X] = \bar{p}$. Given a data set $\boldsymbol{p} = \{p_i \in M\}_{i=1}^N$, an *empirical Fréchet mean* is a minimizer of the quantity

$$\min_{q \in M} \sum_{i=1}^{N} d(q, p_i)^2. \tag{8}$$

The set of empirical Fréchet means is denoted by $\mathbb{E}[\boldsymbol{p}]$.

Given two probability spaces $(\mathcal{X}_i, \mathcal{S}_i, \nu_i)$ for $i = 1, 2$ and a measurable map $F : \mathcal{X}_1 \to \mathcal{X}_2$, we say that the measure ν_2 is the push-forward of the measure ν_1 with respect to F, if $\nu_2(A) = \nu_1(F^{-1}(A))$ for every A in the sigma-algebra $\mathcal{S}_2$. We denote this by $\nu_2 = F_{\#}\nu_1$.

For more about intrinsic statistics on manifolds, see [13].

2.2.2 Geodesic submanifold regression

Geodesic regression on a Riemannian manifold M was introduced by Fletcher [1]. It is a generalization of linear regression, that seeks the geodesic parametrized by $(p, v) \in TM$ that minimizes the quantity

$$E(p, v) = \frac{1}{2} \sum_{i=1}^{N} d(\mathrm{Exp}_p(t_i v), p_i)^2, \tag{9}$$

given the training data $(t_i, p_i) \in \mathbb{R} \times M$ for $i = 1, ..., N$.

This framework has been generalized to deal with more covariates [2]; assume we are given data $(x_i, p_i) \in \mathbb{R}^l \times M$ for $i = 1, ..., N$. Then, we want to solve for the submanifold γ parametrized by $(p, v_1, ..., v_l)$ that minimizes

$$E(p, v_1, ..., v_l) = \frac{1}{2} \sum_{i=1}^{N} d\left(\mathrm{Exp}_p \left(\sum_{j=1}^{l} x_i(j) v_j \right), p_i \right)^2. \tag{10}$$

This is analogous to fitting a hyperplane in the Euclidean case. Another generalization for multiple independent variables was carried out in [3]. Later on in this work, we propose a way to construct priors for the GP regression on manifolds by regressing a geodesic model.

Tangent space geodesic regression is a Naïve generalization of linear regression, achieved by linearizing the space by picking $p \in M$, transforming the data set $(x_i, p_i) \in \mathbb{R}^l \times M$ for $i = 1, ..., N$ into images of the Riemannian logarithmic map at p. Then, one can carry out linear regression in the tangent space and map the result onto the manifold using the exponential map, yielding a quick approximation of geodesic submanifold regression.

3. Wrapped Gaussian processes

We are now ready to introduce *wrapped Gaussian distributions* (WGDs), computing the conditional distribution of two jointly WGD random points on the manifold. This is an essential part of wrapped Gaussian process (WGP) regression on manifolds introduced in the next chapter, alike in the Euclidean case. In this chapter we also introduce WGPs in a formal way, without studying their properties further.

3.1. Wrapped Gaussian distributions

Wrapped Gaussian distributions (WGDs) originated in directional statistics [14]. There exist multiple different ways of generalizing Gaussian distributions to manifolds. For example, Sommer [15] uses an instrinsic, anisotropic diffusion process for the generalization. Pennec [16], on the other hand, generalizes the Gaussian as the distribution maximizing entropy with a fixed mean and covariance. WGDs rely on linearizing the manifold through a wrapping function, in our case the Riemannian exponential map.

Let (M, d) be an n-dimensional Riemannian manifold. We say that a random point X on M follows a *wrapped Gaussian distribution* (WGD), if for some $\mu \in M$ and symmetric positive definite matrix $K \in \mathbb{R}^{n \times n}$

$$X \sim \left(\mathrm{Exp}_\mu\right)_{\#} \left(\mathcal{N}(0, K)\right), \tag{11}$$

denoted by $X \sim \mathcal{N}_M(\mu, K)$. To sample from this distribution, draw v from $\mathcal{N}(0, K)$ and map the sample to the manifold by $\mathrm{Exp}_\mu(v)$. Now, define *the basepoint* and *tangent space covariance* of X as

$$\mu_{\mathcal{N}_M}(X) := \mu, \ \mathrm{Cov}_{\mathcal{N}_M}(X) := K. \tag{12}$$

In the case of infinite injectivity radius $\mu_{\mathcal{N}_M}(X) \in \mathbb{E}[X]$, but not in general [17, Prop. 2.11]. The random points $X_i \sim \mathcal{N}_{M_i}(\mu_i, K_i), i = 1, 2$, are *jointly WGD*, if the random point (X_1, X_2) on $M_1 \times M_2$ is WGD, that is,

$$(X_1, X_2) \sim \mathcal{N}_{M_1 \times M_2}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}K_1 & K_{12}\\ K_{21} & K_2\end{pmatrix}\right), \quad (13)$$

for some matrix $K_{12} = K_{21}^T$.

We now compute the conditional distribution of two jointly WGD random points, which is the core of WGP regression in Section 4.

Theorem 1. *Assume X_1, X_2 are jointly WGD as in (13), then we have the conditional distribution*

$$X_1|(X_2 = p_2) \sim \left(\mathrm{Exp}_{\mu_1}\right)_{\#}\left(\sum_{v\in A}\lambda_v \mathcal{N}(\mu_v, K_v)\right), \quad (14)$$

where

$$\begin{aligned}\mu_v &= K_{12}K_2^{-1}v,\\ K_v &= K_1 - K_{12}K_2^{-1}K_{12}^T,\\ \lambda_v &= \frac{\mathcal{N}(v|\mathbf{0}, K_2)}{\mathbb{P}\{A\}},\\ A &= \{v \in T_{\mu_2}M \mid \mathrm{Exp}_{\mu_2}(v) = p_2\},\\ \mathbb{P}\{A\} &= \sum_{v\in A}\mathcal{N}(v|\mathbf{0}, K_2).\end{aligned} \quad (15)$$

Proof. Pick $p_1 \in M$. Let $B = \mathrm{Exp}_{\mu_1}^{-1}(p_1)$ be the preimage of p_1 in $T_{\mu_1}M$, similarly $A = \mathrm{Exp}_{\mu_2}^{-1}(p_2)$ as above for p_2, and furthermore K be the tangent space covariance of (X_1, X_2) given in (13), then

$$\begin{aligned}&\mathbb{P}\{X_1 = p_1|(X_2 = p_2)\}\\ =&\frac{\mathbb{P}\{u \in B, v \in A\}}{\mathbb{P}\{v \in A\}}\\ =&\sum_{v\in A, u\in B}\frac{\mathcal{N}(v|\mathbf{0}, K_2)}{\mathbb{P}\{A\}}\frac{\mathcal{N}\left((u, v)|\mathbf{0}, K\right)}{\mathcal{N}(v|\mathbf{0}, K_2)}\\ =&\sum_{v\in A, u\in B}\lambda_v\mathcal{N}(u|\mu_v, K_v)\\ =&\mathbb{P}\{Z = p_1\},\end{aligned} \quad (16)$$

where $Z \sim \left(\mathrm{Exp}_{\mu_1}\right)_{\#}\left(\sum_{v\in A}\lambda_v\mathcal{N}(\mu_v, K_v)\right)$, and $\mathcal{N}(u|\mu_v, K_v)$ is the predictive distribution calculated as in the Euclidean case in (3). □

Remark 2. *If the injectivity radius of the exponential map is infinite, then*

$$\begin{aligned}&X_1|(X_2 = p_2)\\ &\sim \left(\mathrm{Exp}_{\mu_1}\right)_{\#}\left(\mathcal{N}\left(\mu_{\mathrm{Log}_{\mu_2}(p_2)}, K_{\mathrm{Log}_{\mu_2}(p_2)}\right)\right),\end{aligned} \quad (17)$$

following the notation in (15). Furthermore, if the probability mass on the area of injectivity of the exponential map is large enough, we can use this expression as a reasonable approximation for the predictive distribution, as the Gaussian mixture distribution in the tangent space can be well approximated by a single Gaussian.

3.2. Wrapped Gaussian processes

A collection f of random points on a manifold M indexed over a set Ω is a *wrapped Gaussian process* (WGP), if every finite subcollection $(f(\omega_i))_{i=1}^N$ is jointly WGD on M^N. We define

$$m(\omega) := \mu_{\mathcal{N}_M}(f(\omega)) \quad (18)$$

$$k(\omega, \omega') := \mathrm{Cov}_{\mathcal{N}_M}(f(\omega), f(\omega')), \quad (19)$$

called the *basepoint function* (BPF) and *tangent space covariance function* (TSCF) of f, respectively. The restriction we have on Ω, is being able to define a kernel on it.

A WGP f can be viewed as a WGD on the possibly infinite-dimensional product manifold $M^{|\Omega|}$. To elaborate, formally one can state

$$f \sim (\mathrm{Exp}_m)_{\#}(\mathcal{GP}(0, k)). \quad (20)$$

The difference is, that the tangent space distribution is a GP instead of a GD. The WGP is entirely characterized by the pair (m, k), similar to the Euclidean case. Therefore, we introduce the notation $f \sim \mathcal{GP}_M(m, k)$.

4. Gaussian process inference on manifolds

In the following, we discuss two different methods of GP regression on a Riemannian manifold M with infinite injectivity radius (or using the approximation in Remark 2), given the noise-free training data

$$\boldsymbol{D}_M = \{(x_i, p_i) \mid x_i \in \mathbb{R}^l,\ p_i \in M,\ i = 1, ..., N\}. \quad (21)$$

For shorthand notation, we denote $\boldsymbol{x} = (x_i)_{i=1}^N$ and $\boldsymbol{p} = (p_i)_{i=1}^N$. Additionally, $\boldsymbol{x}_*$ is used for the test inputs, and $\boldsymbol{p}_*$ for the test outputs. Later, we remark that the first approach is actually a special case of the latter one, see Fig. 2.

4.1. Naïve tangent space approach

Choose $p \in M$ (typically $p \in \mathbb{E}[\boldsymbol{p}]$), and transfrom the training data $\boldsymbol{D}_M$ into $\boldsymbol{D}_{T_pM}$ by

$$\boldsymbol{D}_{T_pM} = (\boldsymbol{x}, \boldsymbol{y}) := \{(x_i, y_i) \mid y_i = \mathrm{Log}_p(p_i)\}, \quad (22)$$

see Fig. 2 a). As $\boldsymbol{D}_{T_pM} \subset \mathbb{R}^l \times T_pM$ now lives in a Euclidean space, fit a GP $f_{euc} \sim \mathcal{GP}(m_{euc}, k_{euc})$ to the data using GP regression, resulting in the predictive distribution $\boldsymbol{y}_*|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*)$. Then, reversing the previous data transformation, we can map the random vector to a random point $\boldsymbol{p}_*|\boldsymbol{p}$ on the manifold M, resulting in

$$\boldsymbol{p}_*|\boldsymbol{p} = \mathrm{Exp}_p(\boldsymbol{y}_*) \sim \left(\mathrm{Exp}_p\right)_{\#}\left(\mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*)\right). \quad (23)$$

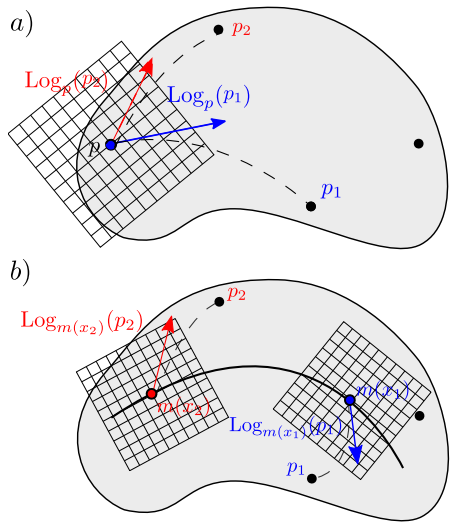

Figure 2. a) Tangent space GP data transformation. Data point p_i (in black) is transformed into $\mathrm{Log}_p(p_i) \in T_pM$. This can be seen as a special case of WGP regression, with a fixed prior BPF $m(x) = p$. In b), the data transformation is visualized with a more general prior BPF m (black curve).

4.2. Wrapped Gaussian process regression

Now we generalize GP regression inside a probabilistic framework, relying on the results presented in Section 3, by assuming a WGP prior $f_{prior} \sim \mathcal{GP}_M(m, k)$. According to the prior, the joint distribution between the training outputs $\boldsymbol{p}$ and test outputs $\boldsymbol{p}_*$ at $\boldsymbol{x}_*$ is given by

$$\begin{pmatrix} \boldsymbol{p}_* \\ \boldsymbol{p} \end{pmatrix} \sim \mathcal{N}_{M_1 \times M_2}\left(\begin{pmatrix} \boldsymbol{m}_* \\ \boldsymbol{m} \end{pmatrix}, \begin{pmatrix} \boldsymbol{k}_{**} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & \boldsymbol{k} \end{pmatrix} \right), \tag{24}$$

where $\boldsymbol{m} = m(\boldsymbol{x})$, $\boldsymbol{m}_* = m(\boldsymbol{x}_*)$, $\boldsymbol{k} = k(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{k}_* = k(\boldsymbol{x}_*, \boldsymbol{x})$, and $\boldsymbol{k}_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$. Therefore, by Theorem 1 and using the approximation in Remark 2 (if necessary)

$$\begin{aligned} \boldsymbol{p}_* | \boldsymbol{p} &\sim \left(\mathrm{Exp}_{\boldsymbol{m}_*}\right)_{\#} \left(\mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*)\right), \\ \boldsymbol{\mu}_* &= \boldsymbol{k}_* \boldsymbol{k}^{-1} \mathrm{Log}_{\boldsymbol{m}} \boldsymbol{p}, \\ \Sigma_* &= \boldsymbol{k}_{**} - \boldsymbol{k}_* \boldsymbol{k}^{-1} \boldsymbol{k}_*^T. \end{aligned} \tag{25}$$

The predictive distribution $\boldsymbol{p}_* | \boldsymbol{p}$ is not necessarily WGD, as $\boldsymbol{\mu}_*$ might be non-zero. The distribution can be sampled from, but computing exactly quantities such as $\mathbb{E}[\boldsymbol{p}_* | \boldsymbol{p}]$ is not trivial. As in [18, Sect. 3.1.1], the distribution can be approximated via Riemannian unscented transform or by using a WGD with the basepoint at $\mathrm{Exp}_{\boldsymbol{m}_*}(\boldsymbol{\mu}_*)$ and parallel transporting the tangent space covariance to this point along the geodesic $\gamma(t) = \mathrm{Exp}_{\boldsymbol{m}_*}(t\boldsymbol{\mu}_*)$.

Remark 3. $\mathrm{Exp}_{\boldsymbol{m}_*}(\boldsymbol{\mu}_*)$ *is not necessarily a Fréchet mean of* $\boldsymbol{p}_* | \boldsymbol{p}$*. However, it is the* maximum a posteriori *(MAP) estimate. For this reason, we will use* $\mathrm{Exp}_{\boldsymbol{m}_*}(\boldsymbol{\mu}_*)$ *as a point prediction in Section 5.*

4.2.1 Choosing a prior

The prior WGP $f_{prior} \sim \mathcal{GP}_M(m, k)$ indexed over Ω is chosen by picking a kernel k on Ω to be the TSCF, and picking a BPF m so that p and $m(x_i)$ live in the same connected component of M for every data-point (x_i, p_i).

In Section 5, two kinds of prior BPFs are used. The first BPF m_1 is a generalization of a centered GP, given by $m_1(\omega) = \bar{p}$, for all $x \in \Omega$ and a $\bar{p} \in \mathbb{E}[\boldsymbol{p}]$. The second kind m_2, uses a previous regression (such as geodesic submanifold regression) γ on the dataset $\boldsymbol{D}_M$. That is, $m_2(\omega) = \gamma(\omega)$ for all $\omega \in \Omega$. For computational reasons, we only consider TSCFs that assume each tangent space coordinate independent, resulting in the *diagonal RBF* kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{diag}(k_1(\boldsymbol{x}, \boldsymbol{x}'), k_2(\boldsymbol{x}, \boldsymbol{x}'), ..., k_n(\boldsymbol{x}, \boldsymbol{x}')), \tag{26}$$

where each k_i are chosen to be RBF kernels, $\mathrm{diag}(A, B)$ is a block-diagonal matrix with blocks A and B, $\boldsymbol{x}, \boldsymbol{x}' \subset \Omega$, and n is the dimension of M. The diagonal RBF yields uncertainty estimates, but not a generative model, as this would need covariance between coordinates.

Optimizing hyperparameters. We choose the TSCF from a parametric family of kernels $\{k_\theta\}_{\theta \in \Theta}$ maximizing the *marginal likelihood*, as in the Euclidean case. In the setting of WGPs, the marginal likelihood becomes

$$\mathbb{P}\{\boldsymbol{p} | \theta\} = \sum_{v \in \mathrm{Exp}_{\boldsymbol{m}}^{-1}(\boldsymbol{p})} \mathcal{N}(v | \boldsymbol{0}, K_\theta), \tag{27}$$

where $K_\theta = k_\theta(\boldsymbol{x}, \boldsymbol{x})$. To improve the approximation discussed in Remark 2, we propose to maximize the quantity

$$\mathbb{P}\{\boldsymbol{p} | \theta\} \approx \mathcal{N}\left(\mathrm{Log}_{\boldsymbol{m}}(\boldsymbol{y}) | \boldsymbol{0}, K_\theta\right), \tag{28}$$

as maximizing this quantity increases the probability mass given by the prior distribution to the area of injectivity. The diagonal RBF kernel (Eq. (26)) can be optimized by choosing each k_i to maximize the marginal likelihood of the respective tangent space coordinate independently. That is, k_i is chosen to maximize the marginal likelihood of the data set $\left\{\left(x_j, \pi_i\left(\mathrm{Log}_{m(x_j)}(p_j)\right)\right)\right\}_{j=1}^N$, where π_i is the projection onto the ith component.

A part of engineering the kernel is to pick a frame for the manifold. A frame is a smooth map $\rho : M \to \mathbb{R}^{n \times n}$, so that the columns of $\rho(p)$ form an orthonormal basis for T_pM. This way, there is a relation between tangent vectors in different tangent spaces, and so the covariance becomes meaningful.

The WGP regression process is summarized in Alg. 4.

Algorithm 4 (WGP regression.)**.** *The following describes step-by-step how to carry out WGP regression.*

Input *Manifold-valued training data* $\boldsymbol{D}_M = \{(x_i, p_i)\}_{i=1}^n$.

Output *Predictive distribution for* $\boldsymbol{p}_*|\boldsymbol{p}$ *at* $\boldsymbol{x}_*$.

i. Choose a prior BPF m.

ii. Transform $\boldsymbol{D}_{T_{\boldsymbol{m}}M} \leftarrow \{(x_i, \mathrm{Log}_{m(x_i)}(p_i))\}_{i=1}^N$.

iii. Choose a prior TSCF k *from a parametric family by optimizing the hyperparameters.*

iv. Using GP prior $\mathcal{GP}(0, k)$*, carry out Euclidean GP regression for the transformed data* $\boldsymbol{D}_{T_{\boldsymbol{m}}M}$*, yielding the mean and covariance* $(\boldsymbol{\mu}_*, \Sigma_*)$.

vi. End with the predictive distribution $\boldsymbol{p}_*|\boldsymbol{p} \sim (\mathrm{Exp}_{\boldsymbol{m}_*})_\#(\mathcal{N}(\boldsymbol{\mu}_*, \Sigma_*))$

4.2.2 Observations with noise

A difficulty arises, when introducing a noise model on our observations. In the Euclidean case, a popular noise model on the observations (x_i, p_i) is given by $p_i = f(x_i) + \epsilon$, where f is the function we approximate and $\epsilon \sim \mathcal{N}(0, K_{\mathrm{err}})$ is the noise term. In [1], this model is generalized to the manifold setting implicitely as

$$p_i = \mathrm{Exp}_{f(x_i)}(\epsilon), \tag{29}$$

which is also supported by the central limit theorem provided in [19]. However, this makes the WGP analytically intractable. To allow computations, we propose the error model $\mathrm{Log}_{m(x_i)}(p_i) = \mathrm{Log}_{m(x_i)}(f(x_i)) + \epsilon$, that is, the error lives in the tangent space of the prior mean at x_i. This can be viewed as a first order approximation of (29) around $m(x_i)$. Introduction of this error changes the regression procedure only slighty; the joint distribution of $\boldsymbol{p}$ and $\boldsymbol{p}_*$ changes into

$$\begin{pmatrix} \boldsymbol{p}_* \\ \boldsymbol{p} \end{pmatrix} \sim \mathcal{N}_{M_1 \times M_2}\left(\begin{pmatrix} \boldsymbol{m}_* \\ \boldsymbol{m} \end{pmatrix}, \begin{pmatrix} \boldsymbol{k}_{**} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^T & \boldsymbol{k} + K_{\mathrm{err}} \end{pmatrix} \right). \tag{30}$$

Rest of the computations are then carried out similarly, with the replacement of $\boldsymbol{k}$ with $\boldsymbol{k} + K_{\mathrm{err}}$ everywhere.

5. Experiments

We demonstrate WGP regression on three manifolds. First, we visualize our algorithm on the 2-sphere using both an illustrative toy dataset and fitting a WGP to motion capture data of the left *femur* of a person walking in a circular pattern. Next, we illustrate DTI upsampling with uncertainty estimates as a tensor field prediction task on a single DTI slice living on the manifold of symmetric and positive definite matrices, and finally we study the effect of age on the shape of Corpus Callosum in Kendall's shape space.

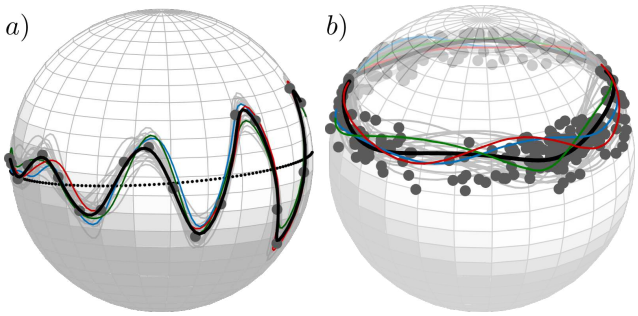

Figure 3. Depicted in a) is WGP regression using a prior BPF given by geodesic regression (dotted black) on a toy data set (grey dots) on S^2. The predictive distribution is visualized using the MAP estimate (black line, see Remark 3) and 20 samples from the distribution (in gray) with three samples emphasized (in red, green and blue). In b), a motion capture dataset of the orientation of the left *femur* of a walking person. The independent variables were estimated by *principal curve* analysis, and a WGP was fitted.

5.1. Data on 2-sphere

As a sanity check, we first visualize our method on a toy dataset on the 2-sphere seen as a Riemannian manifold with the Riemannian metric induced by the Euclidean metric on $\mathbb{R}^3$. This manifold has a finite injectivity radius, thus the approximation presented in Remark 2 is used. A regressed geodesic γ is used as the prior BPF (Sec.3.2), and a diagonal RBF kernel (as in Eq. (26)) with optimized hyperparameters is chosen as the prior TSCF. See Fig. 3 a).

Next, we consider motion capture data of the orientation of the left *femur* of a person walking in a circular pattern [20, 21, 22]. This data naturally lives on S^2 and is periodic. We estimate the periodic independent variables of the data by computing its principal curve as described in [22]. Then, we fit a WGP using Fréchet mean BPF and the TSCF is chosen to be diagonal with the periodic kernel k given by

$$k(t, t') = \sigma^2 \exp\left(-\frac{2\sin^2(|t - t'|/2)}{l^2} \right), \tag{31}$$

where the hyperparameters σ^2 and l^2 are optimized as described in Sect. 4.2.1. Note that the Fréchet mean BPF was used, as the data is not geodesic in trend. The resulting WGP is depicted in Fig. 3 b).

5.2. Diffusion tensor imaging data

We consider a patch of estimated voxel-wise DTI tensors from a coronal slice of an HCP subject [23, 24, 25]. The tensors reside on the manifold $\mathbb{R}^2 \times \mathrm{PD}(3)$, where $\mathrm{PD}(n)$ is the set of $n \times n$ positive definite matrices. When endowed with the affine-invariant metric [26], $\mathrm{PD}(n)$ forms a Riemannian manifold of non-positive curvature, meaning we can perform exact WGP regression with values in $\mathrm{PD}(n)$. The data set consists of 15×19 tensors (elements of $\mathrm{PD}(3)$)

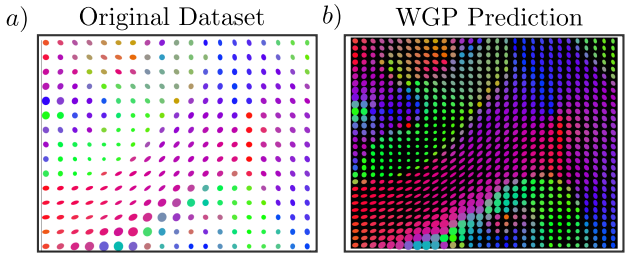

Figure 4. Upsampling DTI tensor field by WGP regression. Colors depict the direction of the principal eigenvector of the respective tensor. a)Tthe slice shown as a tensor field, b) MAP estimate of the predictive distribution of WGP regression on the original data set with uncertainty visualized below (white indicates maximum relative error, black indicates no error). The relative error is computed by dividing by the maximal error over the experiment here and in Fig. 5 c) and e).

with isotropic spacing, see Fig. 4 a). DTI upsampling is performed as an interpolation task on a 30×30 grid, fitting a WGP to the data and estimating up-sampled values using the estimated WGP. As a measure of uncertainty of the result, we calculate the sum of variances of each tangent space coordinate at the interpolated points; this is visualized as a background intensity in Fig. 4 b).

To illustrate the flexibility of WGP regression, we perform a second upsampling experiment, where we randomly subsample only a fifth of the original DTI tensors, see Fig. 5 a). In Fig. 5 c) is shown the corresponding MAP estimate of the predictive distribution (see Remark 3), where empirical Fréchet mean was used as the prior BPF (Fig. 5 b)) and diagonal RBFs with optimized hyperparameters as the prior TSCFs. Finally, to illustrate the effect of the choice of prior BPF, a final experiment used the result of geodesic submanifold regression as the prior BPF, see Fig. 5 d), e).

Note that the tensor field can be reconstructed well even from just 20% of the data, although with increased uncertainty, as can be seen when comparing Figs. 5 c), e) to Fig. 4 a). The predictive WGPs in Figs. 5 c) and e) do not differ vastly, although different BPFs were used. They yield a different result in the upper-left corner area, where the subsampled dataset is not dense, hence the regressed result is influenced by the prior BPF. In the middle, where we also lack information, the resulting tensor fields look similar. The error structures are very similar, seen in Figs. 5c), e). This can be explained by the optimized prior hyperparameters of the TSCFs being similar in both cases (the residuals do not affect the posterior covariance other than through hyperparameter optimization).

5.3. Corpus Callosum data

Next, we turn to a dataset of landmark representations of Corpus Callosum (CC) shapes [1]. A landmark representation is a set of k points in $\mathbb{R}^2$, so that length, translation

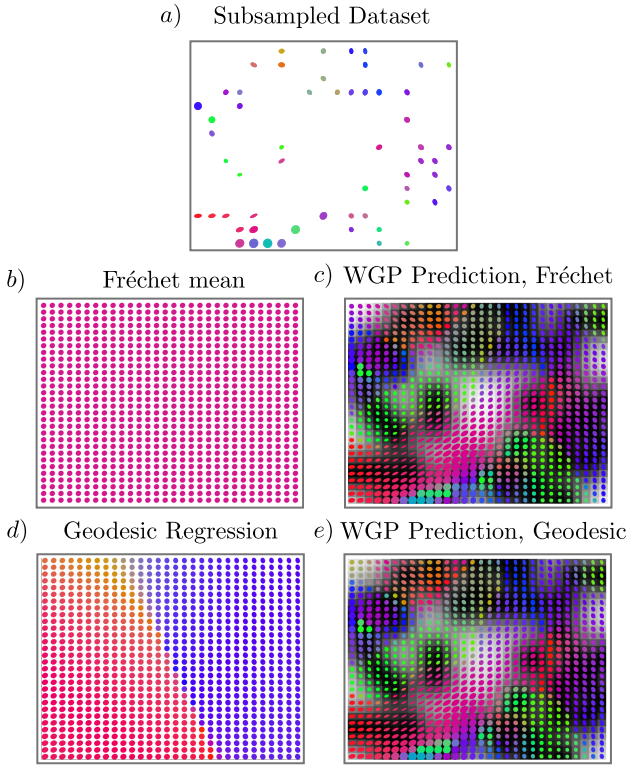

Figure 5. Upsampling DTI tensor field by WGP regression. This time, we carry out the regression on a subsampled tensor field (shown in a)), where only 20% of the elements of the original tensor field (see Fig. 4 a)) are present. We carry out the regression using two different prior WGP BPFs. In b), the first prior BPF using the Fréchet mean is shown and the corresponding predictive WGP is visualized in c), using the MAP estimate to plot the tensors. The second prior BPF is given by geodesic regression, shown in d), with the corresponding predictive WGP in e). For color descriptions, refer to the caption of Fig. 4. The uncertainty fields in c) and e) have similar shapes, but the magnitudes differ.

and rotation factors have been quotiented out, resulting in a point in the *Kendall's shape space* [27]. The dataset consists of 65 shapes, of which we pick randomly 6 to be the test set, the rest are used for training.

Results are presented in Fig. 6. A tangent space geodesic regression is used as the prior BPF, and a diagonal RBF kernel with optimized hyperparameters is used as the prior TSCF. As the CC shapes vary considerably even in the same age group, the WGP predictive mean does not yield notable gains on the tangent space geodesic regression used as prior BPF. However, it provides uncertainty estimates of the shape. Notably, the results imply that aging brings about wider variation in the upper-right part of the CC.

6. Conclusion and discussion

This paper introduced WGP regression on Riemannian manifolds in a novel Bayesian inference framework relying

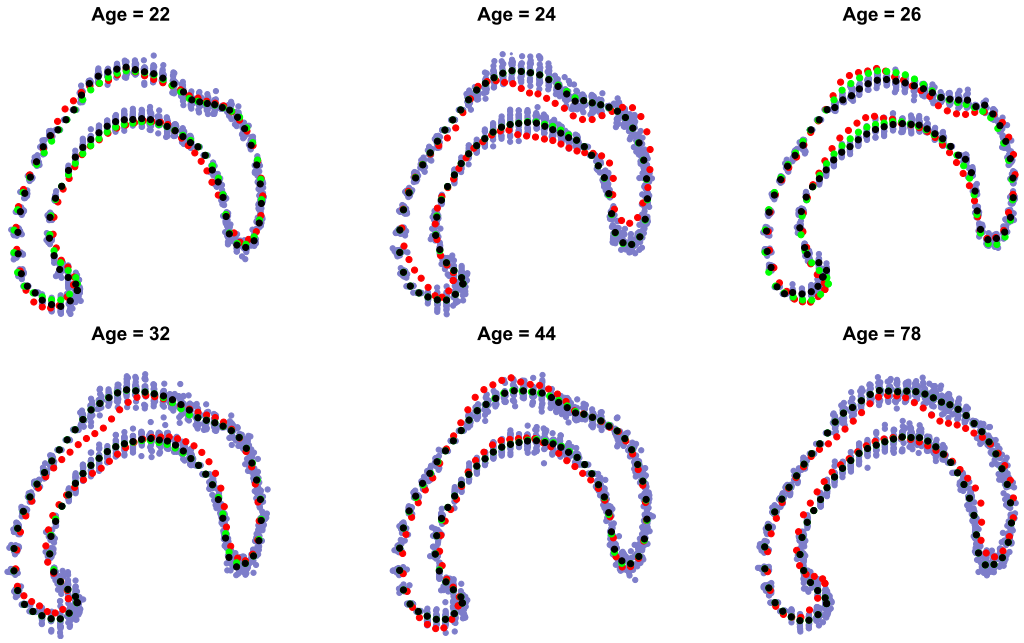

Figure 6. WGP regression applied to a population of Corpus Callosum shapes labeled by age. Red shapes are data points from the test set, not used for training. In black, the MAP estimates of the predictive distributions, in green values of the prior BPF at corresponding ages. Drawn in blue are 20 samples from the predictive distribution.

on WGPs, defined via WGDs. Then, the conditional distribution of two jointly WGD random points was computed for WGP regression. We demonstrated the method on three manifolds; on the 2-sphere using a toy data set and motion capture data of the femur of a walking person, tensor data originating from DTI and on a set of Corpus Callosum shapes. The results of the experiments imply that WGP regression can be used effectively on Riemannian manifolds, providing meaningful uncertainty estimates.

This being the first step, there are still open questions; how to engineer prior distributions efficiently, and how to treat the predictive distribution? The predictive distribution admits an explicit expression, but the prediction is not a WGP anymore. Therefore, we do not have same closure properties of the family of distributions as in the Euclidean case. This leaves open the question, whether one should consider other generalizations of GDs than the wrapped one when carrying out GP regression on manifolds?

We suggested an approximation in Remark 2, not quantifying how reliable it is in the case of non-infinite injectivity radius. In practice the approximation seems plausible (see Fig. 3), but should be studied in more detail. Furthermore, it is of interest, in which cases the computations can be carried out analytically, when the injectivity radius is non-infinite.

The central limit theorem presented in [19] suggests to use WGD distributed error terms, but this poses the difficulty of incorporating the noise term into the prior, when the noise term might live in a different tangent space. The workaround used in this paper was to approximate this error term linearly in the tangent space of the prior BPF, however, other models should also be considered.

Finally, GP regression could be generalized to a broader family of spaces than Riemannian manifolds. In WGP regression, the key is having a wrapping function from a model vector space onto the manifold. For example, another context where such structure appears, is the weak Riemannian structure of the space of probability measures under the Wasserstein metric [28].

7. Acknowledgements

This research was supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation. Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The authors would also like to thank Tom Dela Haije and Søren Hauberg for fruitful discussions and feedback.

References

[1] P. T. Fletcher, "Geodesic regression and the theory of least squares on Riemannian manifolds," *International journal of computer vision*, vol. 105, no. 2, pp. 171–185, 2013. 1, 3, 6, 7

[2] H. J. Kim, N. Adluru, M. D. Collins, M. K. Chung, B. B. Bendlin, S. C. Johnson, R. J. Davidson, and V. Singh, "Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2705–2712, 2014. 1, 3

[3] D. Pigoli, A. Menafoglio, and P. Secchi, "Kriging prediction for manifold-valued random fields," *Journal of Multivariate Analysis*, vol. 145, pp. 117–131, 2016. 1, 2, 3

[4] J. Hinkle, P. Muralidharan, P. T. Fletcher, and S. Joshi, "Polynomial regression on Riemannian manifolds," in *European Conference on Computer Vision*, pp. 1–14, Springer, 2012. 1

[5] M. Banerjee, R. Chakraborty, E. Ofori, D. Vaillancourt, and B. C. Vemuri, "Nonlinear regression on Riemannian manifolds and its applications to Neuro-image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 719–727, Springer, 2015. 1

[6] B. C. Davis, P. T. Fletcher, E. Bullitt, and S. Joshi, "Population shape regression from random design data," *International journal of computer vision*, vol. 90, no. 2, pp. 255–266, 2010. 1

[7] L. Kühnel and S. Sommer, "Stochastic development regression on non-linear manifolds," in *International Conference on Information Processing in Medical Imaging*, pp. 53–64, Springer, 2017. 1

[8] Y. Hong, X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Regression uncertainty on the grassmannian," in *Artificial Intelligence and Statistics*, pp. 785–793, 2017. 1

[9] F. Steinke and M. Hein, "Non-parametric regression between manifolds," in *Advances in Neural Information Processing Systems*, pp. 1561–1568, 2009. 2

[10] G. Jona-Lasinio, A. Gelfand, M. Jona-Lasinio, *et al.*, "Spatial analysis of wave direction data using wrapped gaussian processes," *The Annals of Applied Statistics*, vol. 6, no. 4, pp. 1478–1498, 2012. 2

[11] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, vol. 1. MIT press Cambridge, 2006. 2

[12] M. P. Do Carmo and J. Flaherty Francis, *Riemannian geometry*, vol. 115. Birkhäuser Boston, 1992. 2

[13] X. Pennec, "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements," *Journal of Mathematical Imaging and Vision*, vol. 25, no. 1, pp. 127–154, 2006. 3

[14] K. V. Mardia and P. E. Jupp, *Directional statistics*, vol. 494. John Wiley & Sons, 2009. 3

[15] S. Sommer, "Anisotropic distributions on manifolds: template estimation and most probable paths," in *International Conference on Information Processing in Medical Imaging*, pp. 193–204, Springer, 2015. 3

[16] X. Pennec, "Probabilities and statistics on Riemannian manifolds: Basic tools for geometric measurements.," in *NSIP*, pp. 194–198, 1999. 3

[17] J. Oller, J. M. Corcuera, *et al.*, "Intrinsic analysis of statistical estimation," *The Annals of Statistics*, vol. 23, no. 5, pp. 1562–1581, 1995. 4

[18] S. Hauberg, F. Lauze, and K. S. Pedersen, "Unscented Kalman filtering on Riemannian manifolds," *Journal of mathematical imaging and vision*, vol. 46, no. 1, pp. 103–120, 2013. 5

[19] W. S. Kendall, H. Le, *et al.*, "Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables," *Brazilian Journal of Probability and Statistics*, vol. 25, no. 3, pp. 323–352, 2011. 6, 8

[20] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2014. 6

[21] C. Ionescu, F. Li, and C. Sminchisescu, "Latent structured models for human pose estimation," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2220–2227, IEEE, 2011. 6

[22] S. Hauberg, "Principal curves on riemannian manifolds," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1915–1921, 2016. 6

[23] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, *et al.*, "The minimal preprocessing pipelines for the Human Connectome Project," *Neuroimage*, vol. 80, pp. 105–124, 2013. 6

[24] S. Sotiropoulos, S. Moeller, S. Jbabdi, J. Xu, J. Andersson, E. Auerbach, E. Yacoub, D. Feinberg, K. Setsompop, L. Wald, *et al.*, "Effects of image reconstruction on fiber orientation mapping from multichannel diffusion MRI: reducing the noise floor using SENSE," *Magnetic resonance in medicine*, vol. 70, no. 6, pp. 1682–1689, 2013. 6

[25] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, *et al.*, "The WU-Minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013. 6

[26] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, 2006. 6

[27] D. G. Kendall, "Shape manifolds, procrustean metrics, and complex projective spaces," *Bulletin of the London Mathematical Society*, vol. 16, no. 2, pp. 81–121, 1984. 7

[28] L. Ambrosio and N. Gigli, "A user's guide to optimal transport," in *Modelling and optimisation of flows on networks*, pp. 1–155, Springer, 2013. 8