Bayesian Analysis (2018) **TBA**, Number TBA, pp. 1–20

Extrinsic Gaussian Processes for Regression and Classification on Manifolds

Lizhen Lin*, Niu Mu†, Pokman Cheung‡, and David Dunson§

Abstract. Gaussian processes (GPs) are very widely used for modeling of unknown functions or surfaces in applications ranging from regression to classification to spatial processes. Although there is an increasingly vast literature on applications, methods, theory and algorithms related to GPs, the overwhelming majority of this literature focuses on the case in which the input domain corresponds to a Euclidean space. However, particularly in recent years with the increasing collection of complex data, it is commonly the case that the input domain does not have such a simple form. For example, it is common for the inputs to be restricted to a non-Euclidean manifold, a case which forms the motivation for this article. In particular, we propose a general extrinsic framework for GP modeling on manifolds, which relies on embedding of the manifold into a Euclidean space and then constructing extrinsic kernels for GPs on their images. These extrinsic Gaussian processes (eGPs) are used as prior distributions for unknown functions in Bayesian inferences. Our approach is simple and general, and we show that the eGPs inherit fine theoretical properties from GP models in Euclidean spaces. We consider applications of our models to regression and classification problems with predictors lying in a large class of manifolds, including spheres, planar shape spaces, a space of positive definite matrices, and Grassmannians. Our models can be readily used by practitioners in biological sciences for various regression and classification problems, such as disease diagnosis or detection. Our work is also likely to have impact in spatial statistics when spatial locations are on the sphere or other geometric spaces.

Keywords: extrinsic Gaussian process (eGP), manifold-valued predictors, neuro-imaging, regression on manifold.

1 Introduction

Over the past few decades, Gaussian process (GP) models have emerged as very powerful tools in many problems of statistics and machine learning. In particular, GP models have been widely used in regression and classification, in which a Gaussian process is used as the prior distribution for the regression function or the latent function of a classification map. GP models are particularly appealing in their ability to accurately quantify uncertainty in estimation and prediction. Rasmussen and Williams (2005) provide an overview on GPs in machine learning. van der Vaart and van Zanten (2008,

*Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, Notre Dame, IN, USA, lizhen.lin@nd.edu

†School of Computing, Electronics and Mathematics, Plymouth University, Plymouth, UK, mu.niu@plymouth.ac.uk

‡Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China

§Department of Statistical Science, Duke University, Durham, NC, USA

 https://doi.org/10.1214/18-BA1135

2009) develop theoretical guarantees of GP models in terms of support and posterior asymptotic theory. However, few attempts have been made in developing applicable GP models for *regression and classifications on manifolds* except for some very special cases, such as the 2-dimensional sphere (Hitczenko and Stein, 2012; Guinness and Fuentes, 2016).

One of the paramount challenges in developing GP models on manifolds is constructing *valid covariance kernels*. Castillo et al. (2014) develop an elegant framework for intrinsic GP models on Riemannian manifolds by rescaling solutions of heat equations, but the constructed intrinsic kernels are often impractical to implement. We provide a general and simple solution by first embedding manifolds into Euclidean spaces via *equivariant embeddings*, which are embeddings that preserve a great deal of the geometry of the manifolds, and then *constructing extrinsic kernels* on the image manifold. We refer to the resulting GPs as *extrinsic GPs (eGPs)*. eGPs are shown to inherit appealing properties of GPs defined on Euclidean spaces, and they adapt to the dimension of the manifolds instead of the dimension of the Euclidean space where the manifolds are embedded onto. Another appealing feature of eGPs is their ease of implementation for inference.

One of the motivations for developing GP models on manifolds is the ubiquity of modern data that are represented in various non-conventional forms. In neuroimaging, the diffusion matrices in diffusion tensor imaging (DTI) are 3×3 *positive definite matrices* (Alexander et al., 2007). In engineering and machine learning, pictures or images are often preprocessed or reduced to a collection of *subspaces* (Ho et al., 2004; Teja and Ravi, 2012). In machine vision and medical diagnostics, a digital image can also be represented by a set of k-landmarks, the collection of which form *landmark-based shape spaces* (Kendall, 1984). Other common examples include *orthonormal frames* (Downs et al., 1971), *surfaces*, *curves*, and *networks*(Kolaczyk et al., 2017). Most of the above examples can be described as *manifolds*, which are locally Euclidean spaces with smooth structures.

There are growing needs and practical motivations for studying regression and classification with predictors on known manifolds. For instance, in medical imaging, a common goal is to reliably predict disease status using DTI data or landmark-based digital images. This can be viewed as a classification problem with manifold-valued inputs or predictors. One example is diagnosis of Attention Deficit Hyperactivity Disorder (ADHD) in children based on DTI. There are also many applications in which it is of interest to relate manifold-valued predictors to quantitative traits. One such case is the study of how intelligence quotient relates to the shape contours of certain brain areas (such as the Hippocampus (Bartsch, 2012)). The shape can be represented by a set of landmarks on the boundary of the contours, the collection of which form a shape manifold. Without valid models and appropriate inferential methods for regression and classification on manifolds, making accurate inferences and predictions in the above applications and related settings will remain difficult.

There is already a rich literature on statistical inference for manifold-valued data consisting of i.i.d measurements. Much of this literature focuses on inference on the location and spread of manifold-valued data (Bhattacharya and Patrangenaru, 2003,

2005; Bhattacharya and Lin, 2017). Some model based methods have also been proposed (Bhattacharya and Dunson, 2010b; Lin et al., 2017; Pelletier, 2005). However, regression or classification problems with predictors on manifolds have received much less attention. Bhattacharya and Dunson (2010a) proposed a framework for regression and classification on manifolds by modeling the joint distribution of covariate and response variables (x, y) using a Dirichlet process mixture of product kernels. This joint model induces a nonparametric model for the conditional distribution of y given x with which one can infer the regression/classification function. However, the practical performance of these models is often unsatisfactory as the cluster allocations are driven too much by the marginal distribution of x, a nuisance parameter.

Our work focuses on regression and classification on *known manifolds*. There is, however, an important line of work in manifold learning, where the predictors concentrate around some *unknown lower-dimensional manifold* but are observed in an often higher-dimensional ambient space. The lower-dimensional geometry is often learnt first via dimension reduction tools, based on which a regression model is built (see, e.g., Cheng and Wu (2013)). An interesting exception is due to Yang and Dunson (2016) in which they show that by imposing a Gaussian process prior on the regression function with a covariance kernel defined directly on the ambient space, the posterior distribution yields a posterior contraction rate depending on the intrinsic dimension of the manifold. They assume that the unknown lower-dimensional space where the predictors center around are a class of *submanifolds* of Euclidean space. Many interesting manifolds do not naturally arise as sub-manifolds; in particular, those given as quotient manifolds; projective shape spaces, planar shapes, 3-D shapes, affine shapes and many other manifolds arising as quotient spaces of spheres. Our framework first embeds the manifold onto the Euclidean space via some often non-trivial embeddings and then defines eGPs on the image of the manifolds (including submanifolds as special cases with the embedding given by the identity map).

The paper is organized as follows. Sections 2 introduces eGP models. In section 3, we illustrate the broad utility of eGP models by applying them to a large class of regression/classification problems with predictors lying on various manifolds. Section 4 is devoted to studying the properties of eGP models in terms of mean squared differentiability and posterior contraction rates. Our paper ends with a discussion.

2 Regression and classification on manifolds

Let M be a smooth manifold where the predictors lie. Given data (x_i, y_i) with $x_i \in M$ and $y_i \in \mathbb{R}$ $(i = 1, \ldots, n)$, assume the following regression model

$$y_i = F(x_i) + \epsilon_i, \tag{1}$$

where $F : M \to \mathbb{R}$ is the regression function on M. Here ϵ_i's are some independent errors which determine the likelihood of the regression model. The goal is to develop statistical models for inference on the regression function $F(x)$. The above model can be easily generalized to binary or categorical responses, and $F(x)$ is called the *classification map* in the former case (see (11) for more details on the binary model).

We focus on Bayesian inference on F. Let $\Pi(F)$ be a prior distribution for F, which updates with the data to produce a posterior distribution, based on which inference is carried out. We denote the posterior distribution by $\Pi(F|D)$, where $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is the data. A Gaussian process (GP), which can be viewed as a probability distribution on the space of functions, is one of the most popular candidates for a nonparametric prior for the regression function. The popularity of GP is due to its simple representation, tractability, flexibility for modeling and appealing theoretical properties. We proceed to propose a general extrinsic framework for constructing *GPs on manifolds*.

The usual definition of a GP in a Euclidean space generalizes to a manifold M. A stochastic process $w(x)$ indexed by $x \in M$ is a *Gaussian process on* M if its evaluation at any finite number of points on M follows a multivariate Gaussian distribution. Specifically, we say $w(x)$ is a GP with mean function $\mu(x)$ and covariance kernel $K(\cdot, \cdot)$ if for any $x_1, \ldots, x_n \in M$,

$$\begin{aligned}(w(x_1), \ldots, w(x_n)) \;&\sim\; N\left((\mu(x_1), \ldots, \mu(x_n)), \Sigma\right), \\ &\text{where } \Sigma_{ij} = cov\left(w(x_i), w(x_j)\right) = K(x_i, x_j).\end{aligned}$$

Notice that $K : M \times M \to \mathbb{R}$ is a *positive semi-definite kernel* on M. Namely, for any points $x_1, \ldots, x_n$ on M and real numbers $a_1, \ldots, a_n$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) \geq 0. \tag{2}$$

The fundamental difficulty in imposing a GP prior on a manifold stems from the highly challenging task of constructing a valid covariance kernel $K(\cdot, \cdot)$. Below we describe a simple recipe for constructing valid covariance kernels using an extrinsic approach.

Let $J : M \to \mathbb{R}^D$ be an embedding of M into some higher dimensional Euclidean space $\mathbb{R}^D$ ($D \geq \dim M$) and denote the image of the embedding as $\widetilde{M} = J(M)$. By definition of an embedding, J is a smooth map such that its differential at each point $x \in M$ is an injective map (from the tangent space of M at x to the tangent space of $\mathbb{R}^D$ at $J(x)$), and J is a homeomorphism between M and its image $\widetilde{M}$. Given a positive semi-definite kernel $\widetilde{K}$ on $\mathbb{R}^D$, we can then define a positive semi-definite kernel (and hence the covariance kernel of a GP) on M by

$$K_{ext}(x_1, x_2) = \widetilde{K}(J(x_1), J(x_2)). \tag{3}$$

Indeed, K_{ext} satisfies condition (2) on M because $\widetilde{K}$ satisfies the same condition on $\mathbb{R}^D$, hence in particular on $\widetilde{M} \subset \mathbb{R}^D$. We call the Gaussian process with the covariance kernel $K_{ext}(\cdot, \cdot)$ defined above an *extrinsic Gaussian process (eGP)*.

Remark 1. Note that there are many valid covariance kernel $\widetilde{K}$ available on the Euclidean space in $\mathbb{R}^D$ which allows us to readily construct valid covariance kernels on manifold M via the construction in (3). Depending on the manifolds and applications of interests, both isotropic and non-isotropic kernels for K_{ext} can be constructed by adopting appropriate kernels $\widetilde{K}$ on the image manifold.

We illustrate some popular examples of isotropic kernels. Let $||\cdot||$ be the Euclidean norm. We define the *extrinsic distance* on the manifold M as

$$\rho(x_1, x_2) = \|J(x_1) - J(x_2)\|. \tag{4}$$

One can immediately generalize the popular squared exponential kernel in Euclidean spaces to manifolds by letting

$$K_{ext}(x_1, x_2) = \alpha \exp(-\beta \rho^2(x_1, x_2)), \tag{5}$$

where $\rho(x_1, x_2)$ is the extrinsic distance given in (4). One can also generalize the class of Matérn covariance kernels to manifolds by letting

$$K_{ext}(x_1, x_2) = \sigma^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\sqrt{2\nu}\rho(x_1, x_2)}{\kappa} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\rho(x_1, x_2)}{\kappa} \right), \tag{6}$$

where $\Gamma(\nu)$ is the gamma function, K_ν is the modified Bessel function of the second kind, and κ and ν are non-negative parameters of the covariance. Matérn covariance kernels are often used in spatial statistics with which one can easily control the smoothness of the sample paths with parameter ν. The following is clear.

Proposition 1. *The kernels given in* (5) *and* (6) *are positive semi-definite kernels on* M.

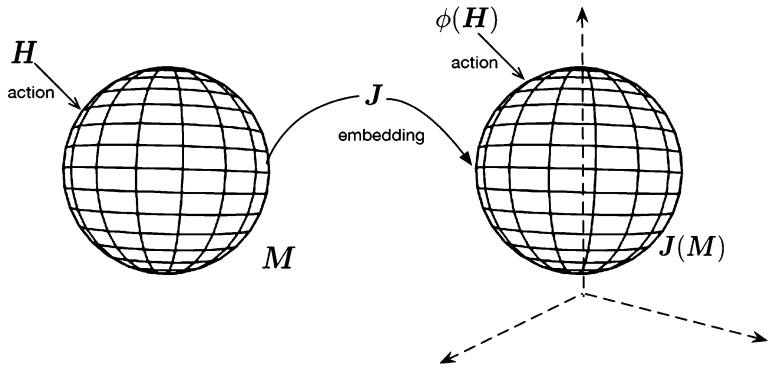

Figure 1: An simple illustration of equivariant embeddings.

Remark 2. The embedding J is never unique. It is desirable to have an embedding that preserves as much geometry as possible. An *equivariant embedding* is one type of embedding that preserves a substantial amount of geometry. Figure 1 provides a visual illustration. Suppose M admits an action of a (usually 'large') Lie group H. Then we say that J is an equivariant embedding if we can find a Lie group homomorphism $\phi : H \to GL(D, \mathbb{R})$ from H to the general linear group $GL(D, \mathbb{R})$ of degree D acting on $\widetilde{M}$ such that

$$J(hp) = \phi(h)J(p)$$

for any $h \in H$ and $p \in M$. The definition seems technical, however, the intuition is clear: if a large group H acts on the manifolds such as by rotation before embedding,

such an action can be preserved via ϕ on the image $\widetilde{M}$. Therefore, the embedding is geometry-preserving in this sense.

Remark 3. The extrinsic method described above has some advantages over using intrinsically defined covariance kernels. In particular, intrinsic kernels are difficult to construct in general. For example, the squared exponential kernel $\alpha \exp(-\beta \rho_g^2(x_1, x_2))$ with ρ_g given by the geodesic or intrinsic distance is in general not a valid kernel. Explicit examples have been found for very special manifolds only, such as spheres. At the same time, simulation tests have shown that there is no significant difference in statistical performance between certain extrinsic and intrinsic models, at least for the example of spheres. However, intrinsic methods are often computationally more complex and expensive.

With a valid covariance kernel on M, one can specify an eGP as a prior $\Pi(F)$ and carry out inference in a Bayesian framework. Given the regression model in (1), we assume that $\epsilon_i \sim N(0, \sigma^2)$, where the parameter σ^2 has a prior distribution such as the inverse gamma distribution with density π_{σ^2}. The prior distribution for the regression function $\Pi(F)$ will be given by the eGP with the covariance kernel in (3). The posterior distribution is given by

$$\Pi\left(U \mid (x_1, y_1), \ldots, (x_n, y_n)\right) = \frac{\int_U \prod_{i=1}^n N(y_i; F(x_i), \sigma^2) \pi_{\sigma^2} \Pi(dF)}{\int \prod_{i=1}^n N(y_i; F(x_i), \sigma^2) \pi_{\sigma^2} \Pi(dF)}, \tag{7}$$

where U is a measurable set in the product space $\mathcal{M} \times (0, \infty)$ with $\mathcal{M}$ denoting the space of all $M \to \mathbb{R}$ regression functions.

Another important class of problems are classification problems, in which one is generally interested in predicting a categorical (e.g., binary as a special case) outcome given the predictors. Denote the responses or outcomes as 1 or 0 for the binary case, and let $F(x)$ be the probability of observing 1 at predictor level x. One can impose a prior distribution on F by imposing an eGP on a latent process $w(x)$, such that $F(x) = L(w(x))$ and L is a fixed link function – for example the probit or logistic link. Properties of $F(x)$ can be derived from those for $w(x)$ as L provides a smooth one-to-one monotone transformation of $w(x)$ into $L(x)$. Extensions to categorical outcomes beyond binary are straightforward.

3 Examples

To illustrate the broad utility of eGP models, we consider a large class of examples with predictors lying on manifolds including spheres, planar shapes, positive definite matrices, and Grassmannians. All details of the embeddings are provided for constructing the extrinsic kernels for eGPs. Embedding manifolds into Euclidean spaces or other manifolds has been applied in different settings. In St. Thomas et al. (2014), for example, the manifold of the parameters of a statistical model is embedded into a big sphere, while Lin et al. (2017) embed the response manifold of a regression model into a Euclidean space for inference. In section 3.1, a simulation study is carried out to compare the performances of an eGP model with that of an intrinsic one in a regression model with

predictors on a sphere. In section 3.2, an eGP model is applied to classify gender of gorillas based on skull images. In this case, the predictor space is the 2-d landmark-based shape space, i.e., the planar shape. In Section 3.3, we consider a classification problem whose predictors are positive definite matrices; this problem has important applications in neuro-imaging. We apply the eGP model to an HIV study in identifying the most sensitive sites for disease detection or diagnostics. Lastly in section 3.4, we apply our eGP model to a regression problem with predictors lying on a Grassmannian manifold in a simulation study.

3.1 Spheres

Modeling on the sphere has received particular attention due to applications in spatial statistics; for example, global models for climate or satellite data (Jun and Stein, 2008; Huang et al., 2011). We consider eGP models for regression with the predictors lying on a sphere S^d. The model is illustrated with predictors on S^2. Note that for the particular case of spheres, there has been a literature investigating valid positive definite functions or covariance functions on the spheres (see. e.g., Gneiting (2013) and Du et al. (2013)).

To construct a valid extrinsic covariance kernel on S^d, first note that S^d is a submanifold of $\mathbb{R}^{d+1}$, so that the inclusion map J serves as a natural embedding of S^d into $\mathbb{R}^{d+1}$. It is easy to check that J is an equivariant embedding with respect to the Lie group $H = SO(d+1)$, the group of $d+1$ by $d+1$ special orthogonal matrices. This embedding preserves the symmetry of the sphere.

One can adopt the extrinsic squared exponential kernel (3) on S^d for an eGP model, with

$$K_{\text{ext}}(x, x') = \alpha \exp\left(-\beta \|J(x) - J(x')\|^2\right) = \alpha \exp\left(-\beta \|x - x'\|^2\right).$$

We now consider a simulation study in which the performance of an eGP model is compared with that of a GP model using an intrinsic kernel. Intrinsic kernels that are computation friendly are only available for some special cases such as S^1 and S^2. We compare eGP to a GP model with the following intrinsic kernel. Letting $d(x, x') = 2 \arcsin\left(\frac{1}{2}\|x - x'\|\right)$, define

$$K_{\text{int}}(x, x') = \alpha \exp\big(-\beta d(x, x')\big), \tag{8}$$

which is a valid covariance kernel on a sphere (e.g., see section 3 of Huang et al. (2011)).

Data are simulated from the regression model,

$$y = F(x_1, x_2, x_3) + \epsilon, \tag{9}$$

where x is a point on the unit sphere, $x_{1:3}$ are the coordinates of x in the three dimensional Euclidean space, the true regression function F is taken to be the sum of $x_{1:3}$ and ϵ is a zero mean Gaussian noise term. We apply a GP model with covariance kernels K_{int} and K_{ext}. Since the kernel parameters ($\theta = \{\alpha, \beta\}$) are correlated (Rasmussen, 2004), standard Markov Chain Monte Carlo (MCMC) sampling traverses the

parameter space slowly. Instead, we use Hamiltonian Monte Carlo (HMC) for inference of kernel parameters which improves efficiency by producing relatively distant proposals that are accepted with high probability (Duane et al., 1987). Here are some details on the priors and the HMC chains: both the length-scale and magnitude hyperparameters of the covariance kernels of the eGP are given gamma(10,10) priors; π_{σ^2} is given by gamma(1,10); the number of Monte Carlo iterations is 10,000 with a burn in of 1,000; The results are not sensitive to varying parameter values of the gamma distributions.

Two kernels are tested using 100 samples with signal-to-noise ratio $26db$. The true function is plotted in red and the estimate is plotted in blue in Figure 2. The horizontal axis is the Euclidean coordinate x_1 and the vertical axis is the functional output. The eGP model appears to produce an estimate that is closer to the true function compare to that from the intrinsic model. Indeed, the eGP model using the kernel K_{ext} yields a smaller root mean square error, which is 0.063 compared to 0.3727 for the intrinsic model. One of the potential reasons for superior performance of eGP over the intrinsic model is non-differentiability of the intrinsic distance hence intrinsic kernel. This non-differentiability can lead to non-smoothness of the Gaussian process (see section 4.1 for more details) thus impacting inference results.

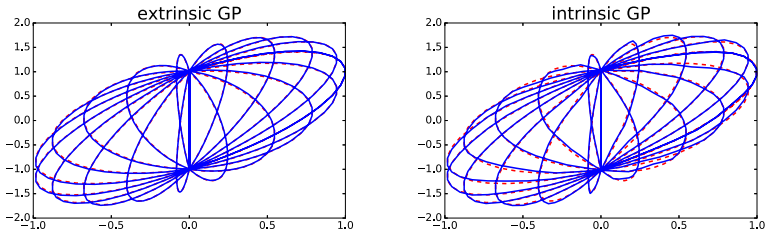

Figure 2: GP predictive results using spherical exponential kernel vs eGP with an extrinsic kernel. Truth is shown in red dashed lines and posterior mean estimates in blue.

3.2 Landmark-based shape spaces Σ_2^k

We now apply eGP models to regression and classification on planar shapes. Planar shape spaces are one of the most important classes of landmark-based shape spaces with wide applications in biology and medical imaging. Such spaces were first studied in Kendall (1977), and in the pioneering work of Bookstein (1978) motivated by applications to biological shapes.

We first describe planar shapes. Let $z = (z_1, \ldots, z_k)$, with $z_1, \ldots, z_k \in \mathbb{R}^2$, be a set of k landmarks. The planar shape Σ_2^k is the collection of zs modulo the Euclidean motions including translation, scaling and rotation. One has $\Sigma_2^k = S^{2k-3}/SO(2)$, the quotient of sphere by the action of $SO(2)$ (or modulo the effect of rotation), the group of 2×2 special orthogonal matrices;

A point in Σ_2^k can be identified as the orbit of some $u \in S^{2k-3}$, which we denote as $\sigma(z)$. Viewing z as a vector of complex numbers, one can embed Σ_2^k into $S(k, \mathbb{C})$, the

space of $k \times k$ complex Hermitian matrices, via the Veronese-Whitney embedding (see e.g. Bhattacharya and Bhattacharya (2012)):

$$J(\sigma(z)) = uu^* = ((u_i \bar{u}_j))_{1 \leq i,j \leq k}. \tag{10}$$

One can verify that J is equivariant (see Kendall (1984)) with respect to the Lie group

$$H = SU(k) = \{A \in GL(k, \mathbb{C}), AA^* = I, \det(A) = I\},$$

with its action on Σ_2^k induced by left multiplication. This embedding J will be used to construct covariance kernels for eGPs on Σ_2^k.

As an example, we apply an eGP to a classification problem with predictors on Σ_2^k. We aim to classify the gorilla skull images from Dryden and Mardia (1998), which are represented as planar shapes with 8 landmarks, by gender. A binary GP classification model is developed using 59 gorilla skull images. We take $y_i \in \{0, 1\}$, where 0 represents a female and 1 a male.

We have the following model:

$$y_i \sim Bernoulli(\pi_i), \quad \pi_i = \Phi(F(x_i)), \quad F(.) \sim \mathrm{GP}(0, K_{ext}), \tag{11}$$

where Φ is the standard normal cdf and K_{ext} is the extrinsic kernel defined in (5).

Following Williams and Rasmussen (1996) and Neal (2012), we used Hamiltonian Monte Carlo (HMC) method for posterior computation. The likelihood is approximated using Laplace's method as in Williams and Barber (1998). Gamma priors are used on the kernel hyperparameters, with gamma(0.5,2) for the length-scale and gamma(50,1) for the magnitude parameter. The number of MCMC iterations is 10,000 with a burn in of 3,000; The HMC estimates of the kernel parameters are shown in Figure 3.

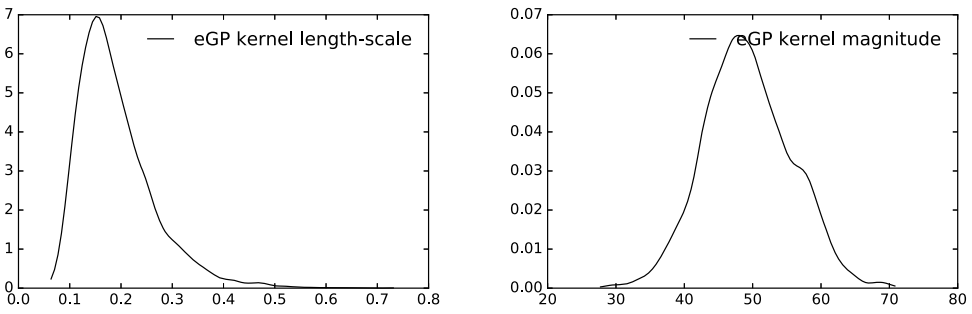

Figure 3: Posterior distributions of the eGP kernel parameters (the length-scale and magnitude).

We use eight skull images as testing data and all these images are successfully classified with our eGP classifier. The classification probabilities are provided in Table 1. The results are compared with a naive GP on the preshape data (modulo the effects of translation and scaling) without any embedding; the latter completely failed at classification by returning all the classification probabilities of 0.5. The results indicate that

Class	female	female	female	female	male	male	male	male
GP classification prob.	7.2e-4	0.319	0.029	0.041	0.96	0.89	0.54	0.86
naive GP classification prob.	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Table 1: Planar shape classification of gender based on gorilla skull shape.

naive GPs are not suitable for complex manifolds not arising as submanifolds of an Euclidean space or when simple representation of the space using Euclidean coordinates is not available. In particular, for complex manifolds such as planar shapes, the naive representation of the data without properly incorporating the underlying geometry (e.g., via equivariant embeddings as in our case), result in a posterior estimate of the latent function that is close to the prior mean (which is zero in our case) thus producing a classification probability of 0.5.

3.3 Diffusion tensor imaging and positive definite matrices

Diffusion tensor imaging (DTI) is designed to measure the diffusion of water molecules in the brain; diffusion tends to be directional along white matter tracks or fibers, corresponding to structural connections between brain regions along which substantial brain activity and communications occur. DTI data are now collected routinely in human studies, and there is abundant interest in using DTI to build better predictive models of cognitive traits and neuropsychiatric disorders. The diffusion anisotropy characterized in terms of *diffusion matrices*, corresponding to 3×3 positive definite matrices measured at each voxel in the brain. We denote the space of all such matrices as SPD(3).

The space SPD(3) belongs to an important class of manifolds that possesses particular geometric structures, which should be taken into account in statistical analyses. Our goal is to study the regression relationship between DTI-valued covariates and patient outcomes.

In order to carry out regression and classification on $SPD(3)$ using our eGP models, we need a nice embedding to construct the extrinsic kernels. There are a few natural embeddings of SPD(3) into Euclidean spaces. In particular, one can embed it into the space Sym(3) of 3×3 real symmetric matrices via the log-map

$$\log : \mathrm{SPD}(3) \to \mathrm{Sym}(3). \tag{12}$$

For $A \in \mathrm{SPD}(3)$ with a spectral decomposition (or diagonalization) $A = U\Lambda U^{-1}$, we have $\log(A) = U \log(\Lambda) U^{-1}$ where $\log(\Lambda)$ is the diagonal matrix whose diagonal entries are the logarithms of the diagonal entries of Λ. The embedding (12) is in fact a diffeomorphism, and is equivariant with respect to the actions of $GL(3, \mathbb{R})$, the 3×3 general linear group, by conjugation. Indeed, for $h \in GL(3, \mathbb{R})$, one has

$$\log(hAh^{-1}) = h \log(A) h^{-1}. \tag{13}$$

Given $A_1, A_2 \in \mathrm{SPD}(3)$, their extrinsic distance under the embedding (12) is given by

$$\rho(A_1, A_2) = \| \log(A_1) - \log(A_2) \|, \tag{14}$$

where $\|\cdot\|$ denotes the Frobenius norm of matrices (i.e. $\|A\| = \mathrm{Tr}(AA^T)^{1/2}$). This extrinsic distance will be used to construct an eGP kernel in (5).

We now consider a diffusion tensor imaging (DTI) data set consisting of 46 subjects with 28 HIV+ subjects and 18 healthy controls. Diffusion tensors were extracted along one atlas fiber tract of the splenium of the corpus callosum. The DTI data for all the subjects are registered in the same atlas space based on arc lengths, with 75 tensors obtained along the fiber tract of each subject. This data set has been studied in a regression setting in Yuan et al. (2012) and in the context of two sample testing (Bhattacharya and Lin (2017)). A GP sampler is carried out between the control group and the HIV+ group for each of the 75 sites along the fiber tract. Therefore, 75 classifiers were run in total. We aim to find out which sites of the splenium of the corpus callosum are most sensitive to influence by HIV.

14 subjects (six controls and eight HIV+) are used to test the HIV status classifiers (0 for healthy and 1 for HIV+) using eGP models. A similar binary GP classification model is applied to the DTI data at each of the prespecified 75 locations along the chosen tract. We have identified the top ten most sensitive sites indexed by the arc length (location on the brain). The results are recorded in Table 2, which shows the total number of correct GP predictions of HIV status of the 14 tested subjects among the top ten sites.

arclength	1.76	4.42	13.56	26.52	31.19	33.16	34.45	35.62	36.80	37.11
# of correct GP prediction	11	11	12	11	11	11	11	11	12	11

Table 2: Diffusion tensor imaging results: top 10 most sensitive sites to influence of HIV.

Again the likelihood is approximated using Laplace approximation techniques and HMC is used for posterior inference. The posterior distribution of kernel hyperparameters for the GP classifier for one of the 75 sites along the fiber tract is shown in Figure 4. Gamma(0.5,2) prior is used for kernel length-scale and gamma(2.5,2) prior for kernel variance. The number of Monte Carlo iterations is 10,000 with a burn in of 3,000.

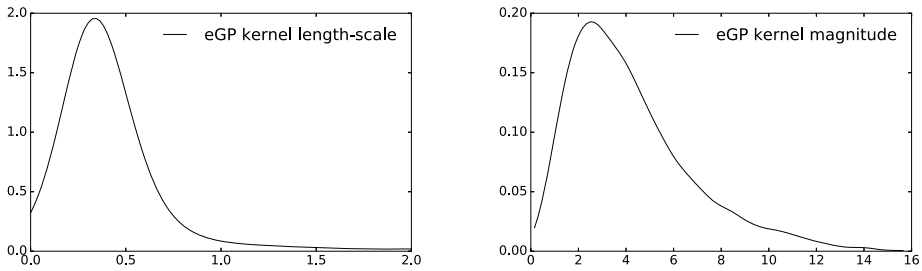

Figure 4: Posterior distribution for the eGP kernel covariance parameters in the diffusion tensor and HIV application.

3.4 Stiefel manifolds and Grassmann manifolds (Grassmannians)

We now consider regression and classification problems whose predictors lie on Stiefel or Grassmann manifolds. Given integers $m \geq k \geq 0$, the Stiefel manifold $V_k(\mathbb{R}^m)$ is the collection of all k-tuples of orthonormal vectors in $\mathbb{R}^m$, and the Grassmann manifold $Gr_k(\mathbb{R}^m)$ is the collection of all k-dimensional subspaces in $\mathbb{R}^m$. Every k-tuple of orthonormal (hence linearly independent) vectors span a k-dimensional subspace, and every k-dimensional subspace is spanned by some k-tuple of orthonormal vectors. This means there is a surjective map $V_k(\mathbb{R}^m) \to Gr_k(\mathbb{R}^m)$. There is a natural action of $O(k)$, the group of $k \times k$ orthogonal matrices, on $V_k(\mathbb{R}^m)$ and any two k-tuples of orthonormal vectors span the same subspace precisely if they differ by an action of $O(k)$, which provides the identification $V_k(\mathbb{R}^m)/O(k) = Gr_k(\mathbb{R}^m)$. Grassmann manifolds have many applications in signal processing and machine learning (Kutyniok et al., 2009).

There is an equivariant embedding of $Gr_k(\mathbb{R}^m)$ into a Euclidean space (Chikuse, 2003). Let $X \in V_k(\mathbb{R}^m)$ and $\sigma(X) = X \cdot O(k)$ be the $O(k)$-orbit of X in $Gr_k(\mathbb{R}^m) = V_k(\mathbb{R}^m)/O(k)$. Note that

$$J(\sigma(X)) = XX'$$

defines an embedding J of $Gr_k(\mathbb{R}^m)$ into the space of $m \times m$ matrices, which may be identified as $\mathbb{R}^{m^2}$. Also, it is equivariant with respect to the group $H = O(m)$ acting on $Gr_k(\mathbb{R}^m)$ via left multiplication on $\mathbb{R}^m$ and on $m \times m$ matrices by conjugation. Indeed, for $h \in H$, one has $J(h\sigma(X)) = hXX'h' = \phi(h)J(\sigma(X))$, where $\phi(h)$ stands for conjugation by h. Now the extrinsic distance between two points in $Gr_k(\mathbb{R}^m)$ is given by

$$\rho(\sigma(X_1), \sigma(X_2)) = \|X_1X_1' - X_2X_2'\|,$$

where $\|\cdot\|$ is the Frobenius norm on matrices. We use the kernel (5).

Remark 4. The Stiefel manifold $V_k(\mathbb{R}^m)$ is naturally a submanifold of $\mathbb{R}^{m\times k}$ and the inclusion map is an equivariant embedding.

We now apply the eGP model to data simulated from $y = F(X) + \epsilon$, where X is an $m \times k$ matrix with $m = 10$ the ambient dimension and $k = 5$ the subspace dimension. The data are simulated from the model with $F(X) = \beta XX'\beta$, where β is some known vector. We simulated 100, 200 and 300 training data points and additional 50 points for testing with different signal-to-noise ratio levels. Table 3 records the RMSE (root mean square error) values. As expected, the RMSE reduces with increasing training size and signal-to-noise ratio.

The posterior distribution of kernel hyperparameters are estimated using HMC. A gamma(2.5,2) prior is used for the kernel length-scale and gamma(40,1) prior for the kernel magnitude. The number of Monte Carlo iterations is 6,000 with a burn in of 1,000.

4 Properties of eGPs

In this section, we first study the properties of an eGP in terms of mean square differentiability. The smoothness of a stochastic process captures and quantifies the intuition

Training size \ Signal-to-noise ratio	10db	20db	30db
$n = 100$	1.25	0.6	0.31
$n = 200$	0.95	0.31	0.098
$n = 300$	0.77	0.27	0.089

Table 3: Simulation results for out-of-sample RMSE for prediction (for 50 testing points) based on predictors on the Grassmannian.

that inputs that are close (on a manifold) are likely to produce similar output values. Therefore, understanding the smoothness property is important for interpolation and prediction. In addition, we show that (see Proposition 4) the posterior contraction rates of eGPs are adaptive to the dimension of the underlying manifold instead of the ambient space where the manifolds are embedded onto building on results from Yang and Dunson (2016).

4.1 Mean square differentiability

We first give the definition of *mean square differentiability and mean square derivative* of a stochastic process on a differentiable manifold. Consider a smooth manifold M and a stochastic process $w(x)$ indexed by $x \in M$. Let $\mu(x)$ and $K(x_1, x_2)$ be the mean and covariance functions of $w(x)$.

Definition 1. (a) Let $x \in M$ and $v \in T_x M$. Choose a smooth path $\gamma : (-\epsilon, \epsilon) \to M$ (for some $\epsilon > 0$) such that $\gamma(0) = x$ and $\gamma'(0) = v$. The stochastic process w is **mean squared (MS) differentiable** *at x with respect to v* if, as $a \to 0$, the random variable

$$\frac{w(\gamma(a)) - w(x)}{a}$$

converges to some limit $D_v w$ in mean squares, i.e.

$$\mathbb{E}\left[\left(\frac{w(\gamma(a)) - w(x)}{a} - D_v w\right)^2\right] \to 0.$$

In this case, $D_v w$ is called the **MS derivative** *of w at x with respect to v*.

(b) If w is MS differentiable at x with respect to every tangent vector at that point, then we simply say that w is *MS differentiable at x*.

(c) If w is MS differentiable at every point in M, then we simply say that w is *MS differentiable (in M)*. In this case, for any tangent vector field V in M, the random variables $\{D_{V_x} w : x \in M\}$ constitute a stochastic process $D_V w$ in M, called the *MS derivative of w with respect to V*.

Remark 5. The definition in (a) depends only on x and v, but otherwise not on the choice of γ. This notion of MS differentiability generalizes the existing one in Euclidean spaces.

Proposition 2. *If the mean function μ is differentiable at x and the covariance function K is of class C^2 at (x,x), then the stochastic process w is MS differentiable at x.*

Proof. Since μ is differentiable at x, the statement will hold for w if it also holds for $w-\mu$, whose mean function is 0. Hence we may assume $\mu = 0$ without loss of generality.

Suppose $\gamma : (-\epsilon, \epsilon) \to M$ is a smooth path with $\gamma(0) = x$ (for some $\epsilon > 0$). Let

$$v = \gamma'(0) \in T_x M, \qquad v^{(1)} = (v, 0),\ v^{(2)} = (0, v) \in T_{(x,x)}(M \times M) = T_x M \times T_x M.$$

For $a \in (-\epsilon, 0) \cup (0, \epsilon)$, consider the random variable

$$D_a = \frac{w(\gamma(a)) - w(x)}{a}.$$

It suffices to show that D_a has a limit in mean squares (i.e. in L^2) as $a \to 0$. Notice that

$$\mathbb{E}[D_a D_b] = \frac{1}{ab}\Big(K(\gamma(a),\gamma(b)) - K(\gamma(a),x) - K(x,\gamma(b)) + K(x,x)\Big).$$

Since K is of class C^2 at (x,x), as $(a,b) \to (0,0)$, we have

$$\mathbb{E}[D_a D_b] \to \big(D_{v^{(1)}} D_{v^{(2)}} K\big)(x,x).$$

It follows that, under the same limit,

$$\begin{aligned}\mathbb{E}[(D_a - D_b)^2] &= \mathbb{E}[D_a^2] + \mathbb{E}[D_b^2] - 2\mathbb{E}[D_a D_b] \\ &\to \big(D_{v^{(1)}} D_{v^{(2)}} K\big)(x,x) + \big(D_{v^{(1)}} D_{v^{(2)}} K\big)(x,x) - 2\big(D_{v^{(1)}} D_{v^{(2)}} K\big)(x,x) \\ &= 0.\end{aligned}$$

Therefore, as $a \to 0$, D_a satisfies the Cauchy condition with respect to the L^2 norm and, by completeness, admits an L^2 limit. □

Proposition 3. *If the mean function μ is differentiable in M and the covariance function K is of class C^2 in $M \times M$, then the stochastic process w is MS differentiable in M. In this case, for any tangent vector field V in M, the MS derivative $D_V w$ has mean function $D_V \mu$ and covariance function $D_{V^{(1)}} D_{V^{(2)}} K$, where $V^{(1)}$ and $V^{(2)}$ are the tangent vector fields in $M \times M$ with $V^{(1)}_{(x_1,x_2)} = (V_{x_1}, 0)$ and $V^{(2)}_{(x_1,x_2)} = (0, V_{x_2})$.*

Proof. The first statement is immediate from Proposition 2. For $i = 1, 2$, let $x_i \in M$ and $\gamma_i : (-\epsilon, \epsilon) \to M$ be a smooth path with $\gamma_i(0) = x_i$ and $\gamma_i'(0) = V_{x_i}$. By the Cauchy-Schwarz inequality and the MS differentiability of w, we have

$$\begin{aligned} & \mathbb{E}\left[\left((D_V w)(x_1) - \frac{w(\gamma_1(a)) - w(x_1)}{a}\right)\right] \to 0, \quad \text{as } a \to 0 \\ \iff \quad & \mathbb{E}[(D_V w)(x_1)] - \frac{\mu(\gamma_1(a)) - \mu(x_1)}{a} \to 0, \quad \text{as } a \to 0 \end{aligned}$$

so that $\mathbb{E}[(D_V w)(x_1)] = (D_V \mu)(x_1)$. Now let $\tilde{w} = w - \mu$. Similarly as above, we have

$$\mathbb{E}\left[\left((D_V \tilde{w})(x_1) - \frac{\tilde{w}(\gamma_1(a)) - \tilde{w}(x_1)}{a}\right)\tilde{w}(x_2)\right] \to 0, \quad \text{as } a \to 0$$

$$\iff \quad \mathbb{E}[(D_V \tilde{w})(x_1)\,\tilde{w}(x_2)] - \frac{K(\gamma_1(a), x_2) - K(x_1, x_2)}{a} \to 0, \quad \text{as } a \to 0$$

so that $\mathbb{E}[(D_V \tilde{w})(x_1)\,\tilde{w}(x_2)] = (D_{V^{(1)}} K)(x_1, x_2)$. Similarly again, we also have

$$\mathbb{E}\left[\left((D_V \tilde{w})(x_1) - \frac{\tilde{w}(\gamma_1(a)) - \tilde{w}(x_1)}{a}\right)\left((D_V \tilde{w})(x_2) - \frac{\tilde{w}(\gamma_2(b)) - \tilde{w}(x_2)}{b}\right)\right] \to 0$$

$$\iff \quad \mathbb{E}[(D_V \tilde{w})(x_1)\,(D_V \tilde{w})(x_2)] - \frac{K(\gamma_1(a), \gamma_2(b)) - K(\gamma_1(a), x_2) - K(x_1, \gamma_2(b)) + K(x_1, x_2)}{ab} \to 0$$

as $(a, b) \to (0, 0)$, which means

$$\begin{aligned}
&\mathbb{E}[(D_V \tilde{w})(x_1)\,(D_V \tilde{w})(x_2)] \\
&\quad = (D_{V^{(2)}} D_{V^{(1)}} K)(x_1, x_2) + (D_{V^{(1)}} D_{V^{(2)}} K)(x_1, x_2) - (D_{V^{(1)}} D_{V^{(2)}} K)(x_1, x_2) \\
&\quad = (D_{V^{(1)}} D_{V^{(2)}} K)(x_1, x_2).
\end{aligned}$$

This completes the proof. □

Corollary 1. *If μ is of class C^n and K is of class C^{2n}, then w is n-times MS differentiable.*

Proof. Repeatedly apply Proposition 3. □

Example 1. Suppose $J : M \to \mathbb{R}^D$ is an embedding of M into a (higher-dimensional) Euclidean space $\mathbb{R}^D$. Given a stochastic process w in $\mathbb{R}^D$, we can pull it back to a stochastic process $J^* w$ in M, with

$$(J^* w)(x) = w(J(x)), \quad \text{for } x \in M.$$

Clearly, if the mean and covariance functions of w are μ and K, then the mean and covariance functions of $J^* f$ are $J^* \mu$ and $(J \times J)^* K$. Also, if μ is C^n, K is C^{2n} and J is C^{2n} as well, then $J^* \mu$ is C^n and $(J \times J)^* K$ is C^{2n}; and hence by Corollary 1, $J^* w$ is n-times MS differentiable.

For example, if w is a Gaussian process in $\mathbb{R}^D$ with a Matérn-ν covariance function (and zero mean), then $J^* w$ is an $\lfloor \frac{\nu - 1}{2} \rfloor$-times MS differentiable Gaussian process in M; and if w is a Gaussian process in $\mathbb{R}^D$ with a squared-exponential covariance function, then $J^* w$ is an infinitely MS differentiable Gaussian process in M.

4.2 Posterior contraction rates of eGPs

In this short subsection, we explore the posterior contraction rates of a regression model on a manifold with eGP as the prior for the regression function. Posterior contrac-

tion rates measure how fast the posterior concentrates in small neighborhoods of the true regression function, providing frequentist asymptotic guarantees on the behavior of the eGP posterior. Given data (x_i, y_i) with $x_i \in M$ and $y_i \in \mathbb{R}$ $(i = 1, \ldots, n)$, assume the regression model (1) where $y_i = F(x_i) + \epsilon_i$, $x_i \in M$ and $\epsilon_i \sim N(0, \sigma^2)$. The prior distribution $\Pi(F)$ will be given by the eGP with the covariance kernel (5) (with a fixed magnitude). The length-scale parameter β is assumed a prior π_β such that β^d follows a gamma distribution gamma(a_0, b_0), where d is the dimension of manifold. For simplicity in exposition, assume σ is known though the results are straightforward to generalize to unknown σ. The posterior distribution of F is then given by

$$\Pi\left(U \mid (x_1, y_1), \ldots, (x_n, y_n)\right) = \frac{\int_U \prod_{i=1}^n f(y_i; F(x_i), \sigma^2)\Pi(dF)}{\int \prod_{i=1}^n f(y_i; F(x_i), \sigma^2)\Pi(dF)}, \tag{15}$$

where U is a measurable set in the space of regression functions and $f(y_i; F(x_i), \sigma^2)$ is the value of a normal density (with mean $F(x_i)$ and variance σ^2) evaluated at y_i. Let F_0 be the true regression function. We say the eGP posterior *contracts to* F_0 *at a rate of* ϵ_n if

$$\Pi\left(U_{\epsilon_n}(F_0)^C \mid (x_1, y_1), \ldots, (x_n, y_n)\right) \to 0, \ a.s. P_{F_0}^n, \tag{16}$$

where $U_{\epsilon_n}(F_0)^C = \{F : d_{\mathcal{M}}(F, F_0) > R\epsilon_n\}$, as $n \to \infty$ for some large constant R and distance $d_{\mathcal{M}}$. We have the following proposition.

Proposition 4. *Assume the regression model (1) with an eGP prior with covariance kernel (5), the following holds.*

(a) Assume M is a smooth and compact manifold and the covariates are from a fixed design. Let $F_0 \in C^s(M)$ $(s \leq 2)$, the s-Hölder smooth class of functions on M, then the posterior distribution of eGP contracts to the true regression function F_0 at a rate of $\epsilon_n = n^{-s/(2s+d)}(\log n)^{d+1}$ with $d_{\mathcal{M}}(F, F_0) = \frac{1}{n}\sum_{i=1}^n |F(x_i) - F_0(x_i)|$.

(b) Assume M is a smooth and compact manifold and the covariates are from a random design with $x_i \sim G(\cdot)$, $i = 1, \ldots, n$, for some distribution $G(\cdot)$ on M with density $g(x)$. Then the results in part (a) hold with $U_{\epsilon_n}(F_0)^C = \{F : \int_{x \in M}(F_A(x) - F_0(x))^2 g(dx) > R\epsilon_n\}$, where $F_A(x) = (F \vee (-A)) \wedge A$, for some A large enough.

Proof. (a) Given the embedding $J : M \to \mathbb{R}^D$, $\tilde{M} = J(M)$ is a d-dimensional submanifold of $\mathbb{R}^D$. Any function $F \in \mathcal{M}$ on M induces a function $\tilde{F} = F \circ J^{-1}$ on $\tilde{M}$. One has

$$y_i = \tilde{F}(\tilde{x}_i) + \epsilon_i,$$

where $\tilde{x}_i = J(x_i) \in \tilde{M}$. Then by Theorem 2.1 of Yang and Dunson (2016), one has

$$\Pi\left(\tilde{U}_{\epsilon_n}(\tilde{F}_0)^C \mid (\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n)\right) \to 0,$$

where $\tilde{U}_{\epsilon_n}(F_0) = \{\tilde{F} : \frac{1}{n}\sum_{i=1}^n \mid \tilde{F}(\tilde{x}_i) - \tilde{F}_0(\tilde{x}_i) \mid < R\epsilon_n\}$. There is a one-to-one correspondence (a bijection) between $\tilde{F}$ and F, and one has $U_{\epsilon_n}(F_0) = \{F : \frac{1}{n}\sum_{i=1}^n |F(x_i) - F_0(x_i)| = \frac{1}{n}\sum_{i=1}^n \mid \tilde{F}(\tilde{x}_i) - \tilde{F}_0(\tilde{x}_i) \mid < R\epsilon_n\}$. Then

$$\Pi\left(U_{\epsilon_n}(F_0)^C \mid (x_1, y_1), \ldots, (x_n, y_n)\right) \to 0,$$

where ϵ_n is given in part (a).

(b) Similar proofs follow from part (a) noting that there is one-to-one correspondence between $\{\tilde{F} : \int_{\tilde{M}} (\tilde{F}(\tilde{x}) - \tilde{F}_0(\tilde{x}))^2 \tilde{g}(\tilde{x}) d\tilde{x} < \epsilon_n\}$ and $\{F : \int_M (F(x) - F_0(x))^2 g(x) dx < \epsilon_n\}$, where $\tilde{g}(\tilde{x})$ is the density on $\tilde{M}$ induced by the embedding J and the density $g(x)$ on M. □

5 Discussion and conclusion

We propose a general extrinsic framework for constructing Gaussian processes on manifolds for regression and classification with manifold-valued predictors. Such models are general, easy to implement and shown to inherit good properties from Gaussian processes on Euclidean spaces. Applications are considered by applying eGP models to regression and classification problems with predictors on a large class of manifolds ranging from spheres, landmark-based shapes spaces, to the spaces of positive definite matrices and Grassmannians. Our work will likely help practitioners make more accurate predictions or diagnoses based on medical imaging. Although the work focuses on regression and classification, eGPs can be used in much broader settings such as in exponential family models for the response y_i given x_i. In addition, eGPs can be certainly used for spatial modeling where the spatial domain is some geometric space such as the sphere. Future work will be devoted to constructing applicable covariance kernels employing the intrinsic Riemannian geometry of manifolds, which are only currently available for a very limited class of manifolds, and also constructing valid GP models for spaces beyond manifolds such as stratified spaces of interests.

References

Alexander, A., Lee, J. E., Lazar, M., and Field, A. S. (2007). "Diffusion Tensor Imaging of the Brain." *Neurotherapeutics*, 4(3): 316–329. 2

Bartsch, T. (2012). *The Clinical Neurobiology of the Hippocampus: An Integrative View*. OUP Oxford. 2

Bhattacharya, A. and Bhattacharya, R. (2012). *Nonparametric Inference on Manifolds: with Applications to Shape Spaces*. Cambridge University Press. IMS monographs #2. MR2934285. doi: https://doi.org/10.1017/CBO9781139094764. 9

Bhattacharya, A. and Dunson, D. (2010a). "Nonparametric Bayes regression and classification through mixtures of product kernels." *Bayesian Analysis*, 9: 145–164. MR3204005. doi: https://doi.org/10.1093/acprof:oso/9780199694587.003.0005. 3

Bhattacharya, A. and Dunson, D. B. (2010b). "Nonparametric Bayesian density estimation on manifolds with applications to planar shapes." *Biometrika*, 97(4): 851–865. MR2746156. doi: https://doi.org/10.1093/biomet/asq044. 3

Bhattacharya, R. and Lin, L. (2017). "Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces." *The Proceedings of the American Mathematical Society*, 145: 413–428. MR3565392. doi: https://doi.org/10.1090/proc/13216. 2, 11

Bhattacharya, R. and Patrangenaru, V. (2005). "Large sample theory of intrinsic and extrinsic sample means on manifolds. II." *The Annals of Statistics*, 33(3): 1225–1259. MR2195634. doi: https://doi.org/10.1214/009053605000000093. 2

Bhattacharya, R. N. and Patrangenaru, V. (2003). "Large sample theory of intrinsic and extrinsic sample means on manifolds." *The Annals of Statistics*, 31: 1–29. MR1962498. doi: https://doi.org/10.1214/aos/1046294456. 2

Bookstein, F. (1978). *The Measurement of Biological Shape and Shape Change*. Lecture Notes in Biomathematics, Springer, Berlin. 8

Castillo, I., Kerkyacharian, G., and Picard, D. (2014). "Thomas Bayes's walk on manifolds." *Probability Theory and Related Fields*, 158(3–4): 665–710. MR3176362. doi: https://doi.org/10.1007/s00440-013-0493-0. 2

Cheng, M. and Wu, H. (2013). "Local Linear Regression on Manifolds and Its Geometric Interpretation." *Journal of the American Statistical Association*, 108(504): 1421–1434. MR3174718. doi: https://doi.org/10.1080/01621459.2013.827984. 3

Chikuse, Y. (2003). *Statistics on Special Manifolds*. Springer, New York. MR1960435. doi: https://doi.org/10.1007/978-0-387-21540-2. 12

Downs, T., Liebman, J., and Mackay, W. (1971). "Statistical methods for vectorcardiogram orientations." *In Vectorcardiography 2: Proc. XIth International Symposium on Vectorcardiography*, 216–222. North-Holland, Amsterdam. 2

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. Wiley, New York. MR1646114. 9

Du, J., Ma, C., and Li, Y. (2013). "Isotropic Variogram Matrix Functions on Spheres." *Mathematical Geosciences*, 45(3): 341–357. MR3107159. doi: https://doi.org/10.1007/s11004-013-9441-x. 7

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). "Hybrid Monte Carlo." *Physics letters B*, 195(2): 216–222. 8

Gneiting, T. (2013). "Strictly and non-strictly positive definite functions on spheres." *Bernoulli*, 19(4): 1327–1349. MR3102554. doi: https://doi.org/10.3150/12-BEJSP06. 7

Guinness, J. and Fuentes, M. (2016). "Isotropic covariance functions on spheres: Some properties and modeling considerations." *Journal of Multivariate Analysis*, 143: 143–152. MR3431424. doi: https://doi.org/10.1016/j.jmva.2015.08.018. 2

Hitczenko, M. and Stein, M. (2012). "Some theory for anisotropic processes on the sphere." *Statistical Methodology*, 9(1–2): 211–227. Special Issue on Astrostatistics + Special Issue on Spatial Statistics. MR2863609. doi: https://doi.org/10.1016/j.stamet.2011.01.010. 2

Ho, J., Lee, K.-C., Yang, M.-H., and Kriegman, D. (2004). "Visual tracking using learned linear subspaces." In *CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, 782–789. 2

Huang, C., Zhang, H., and Robeson, S. (2011). "On the Validity of Commonly Used Covariance and Variogram Functions on the Sphere." *Mathematical Geosciences*, 43(6): 721–733. MR2824128. doi: https://doi.org/10.1007/s11004-011-9344-7. 7

Jun, M. and Stein, M. L. (2008). "Nonstationary covariance models for global data." *The Annals of Applied Statistics*, 2(4): 1271–1289. MR2655659. doi: https://doi.org/10.1214/08-AOAS183. 7

Kendall, D. G. (1977). "The diffusion of shape." *Advances in Applied Probability*, 9: 428–430. 8

Kendall, D. G. (1984). "Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces." *Bulletin of the London Mathematical Society*, 16: 81–121. MR0737237. doi: https://doi.org/10.1112/blms/16.2.81. 2, 9

Kolaczyk, E., Lin, L., Rosenberg, S., and Walters, J. (2017). "Averages of Unlabeled Networks: Geometric Characterization and Asymptotic Behavior." *ArXiv e-prints*. 2

Kutyniok, G., Pezeshki, A., Calderbank, R., and Liu, T. (2009). "Robust dimension reduction, fusion frames, and Grassmannian packings." *Applied and Computational Harmonic Analysis*, 26(1): 64–76. MR2467935. doi: https://doi.org/10.1016/j.acha.2008.03.001. 12

Lin, L., Rao, V., and Dunson, D. B. (2017). "Bayesian nonparametric inference on the Stiefel manifold." *Statistics Sinica*, 27: 535–553. MR3674685. 3

Lin, L., Thomas, B. S., Zhu, H., and Dunson, D. B. (2017). "Extrinsic Local Regression on Manifold-Valued Data." *Journal of the American Statistical Association*, 112(519): 1261–1273. MR3735375. doi: https://doi.org/10.1080/01621459.2016.1208615. 6

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media. 9

Pelletier, B. (2005). "Kernel density estimation on Riemannian manifolds." *Statistics and Probability Letters*, 73(3): 297–304. MR2179289. doi: https://doi.org/10.1016/j.spl.2005.04.004. 3

Rasmussen, C. E. (2004). "Gaussian Processes in Machine Learning." *Advanced Lectures on Machine Learning*, 63–71. 7

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press. MR2514435. 1

St. Thomas, B., Lin, L., Lim, L.-H., and Mukherjee, S. (2014). "Learning subspaces of different dimension." *ArXiv e-prints*, 1404.6841. 6

Teja, G. and Ravi, S. (2012). "Face recognition using subspaces techniques." In *Recent Trends In Information Technology (ICRTIT), 2012 International Conference on*, 103–107. 2

van der Vaart, A. W. and van Zanten, J. H. (2008). "Rates of contraction of posterior distributions based on Gaussian process priors." *The Annals of Statistics*, 36(3): 1435–1463. MR2418663. doi: https://doi.org/10.1214/009053607000000613. 1

van der Vaart, A. W. and van Zanten, J. H. (2009). "Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth." *The Annals of Statistics*, 37(5B): 2655–2675. MR2541442. doi: https://doi.org/10.1214/08-AOS678. 1

Williams, C. K. and Barber, D. (1998). "Bayesian classification with Gaussian processes." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12): 1342–1351. 9

Williams, C. K. and Rasmussen, C. E. (1996). "Gaussian processes for regression." *Advances in neural information processing systems 8*, 514–520. 9

Yang, Y. and Dunson, D. B. (2016). "Bayesian manifold regression." *The Annals of Statistics*, 44(2): 876–905. MR3476620. doi: https://doi.org/10.1214/15-AOS1390. 3, 13, 16

Yuan, Y., Zhu, H., Lin, W., and Marron, J. S. (2012). "Local polynomial regression for symmetric positive definite matrices." *Journal of the Royal Statistical Society: Series B*, 74: 697–719. MR2965956. doi: https://doi.org/10.1111/j.1467-9868.2011.01022.x. 11

Acknowledgments

We are grateful to the Associate Editor and the referees for their helpful suggestions. We thank Professor Hongtu Zhu for providing us the diffusion tensor imaging data used in Section 3. Lizhen Lin acknowledges the support of NSF grants IIS1663870 and DMS CAREER 1654579, and a DARPA grant N66001-17-1-4041.