ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)
81, *Part* 3, *pp.* 603–627

Intrinsic Gaussian processes on complex constrained domains

Mu Niu,
Plymouth University, UK

Pokman Cheung,
London, UK

Lizhen Lin,
University of Notre Dame, USA

Zhenwen Dai and Neil Lawrence
University of Sheffield and Amazon, UK

and David Dunson
Duke University, Durham, USA

[Received January 2018. Final revision February 2019]

Summary. We propose a class of intrinsic Gaussian processes (GPs) for interpolation, regression and classification on manifolds with a primary focus on complex constrained domains or irregularly shaped spaces arising as subsets or submanifolds of $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{R}^3$ and beyond. For example, intrinsic GPs can accommodate spatial domains arising as complex subsets of Euclidean space. Intrinsic GPs respect the potentially complex boundary or interior conditions as well as the intrinsic geometry of the spaces. The key novelty of the approach proposed is to utilize the relationship between heat kernels and the transition density of Brownian motion on manifolds for constructing and approximating valid and computationally feasible covariance kernels. This enables intrinsic GPs to be practically applied in great generality, whereas existing approaches for smoothing on constrained domains are limited to simple special cases. The broad utilities of the intrinsic GP approach are illustrated through simulation studies and data examples.

Keywords: Brownian motion; Constrained domain; Gaussian process; Heat kernel; Intrinsic covariance kernel; Manifold

1. Introduction

In recent years it has become commonplace to collect data that are restricted to a complex constrained space. For example, data may be collected in a spatial domain but restricted to a complex or intricately structured region corresponding to a geographic feature, such as a lake. To illustrate, refer to Fig. 1(b), which plots satellite measurements on chlorophyll levels in the Aral sea (Wood *et al*., 2008). In building a spatial map of chlorophyll levels in this sea, and in conducting corresponding inferences and prediction tasks, it is important to take into account

Address for correspondence: Lizhen Lin, Department of Applied and Computational Mathematics and Statistics, 153 Hurley Hall, University of Notre Dame, Notre Dame, IN 46556, USA.
E-mail: lizhen.lin@nd.edu

 1369–7412/19/81603

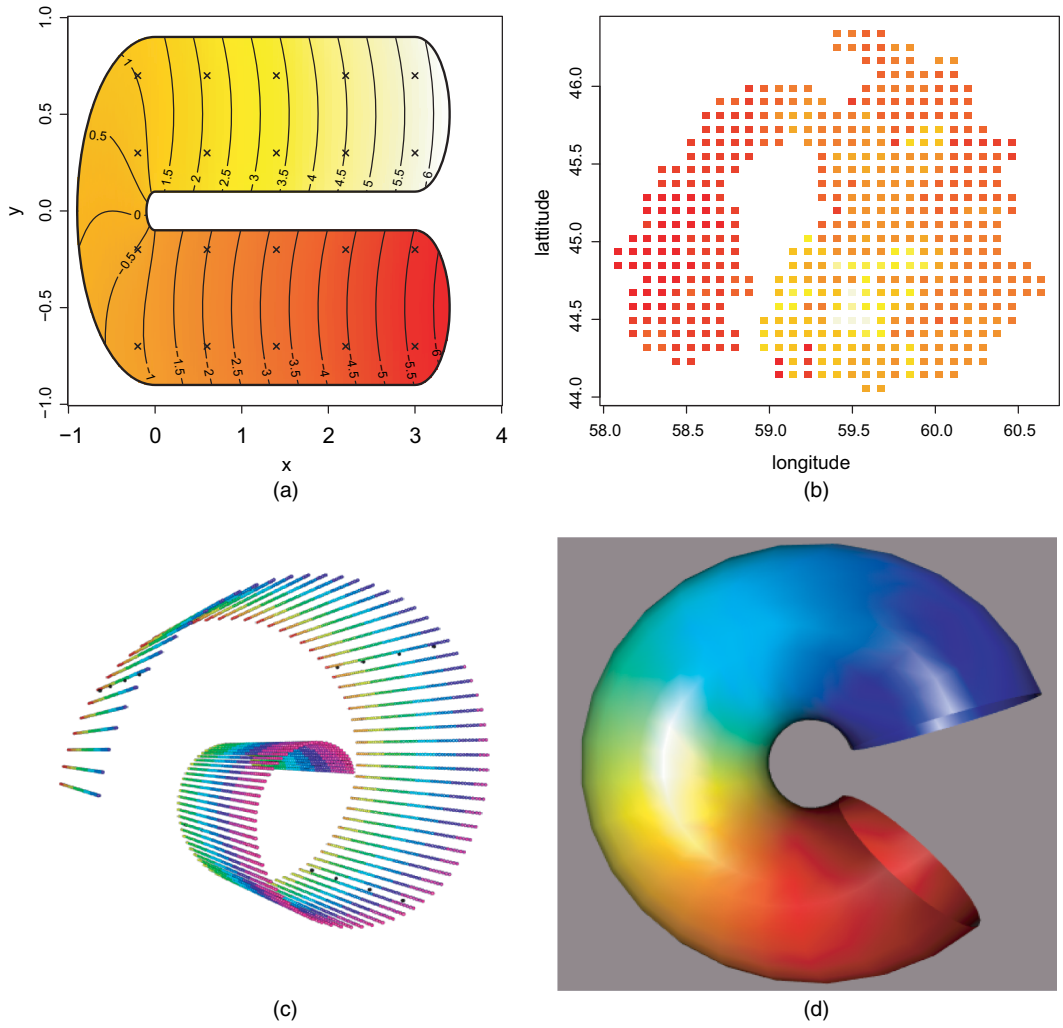

Fig. 1. Illustrative examples: in (a), a test function increases smoothly from the lower right to the upper right within the U-shaped boundary; remote sensed chlorophyll data in the Aral sea from the SeaWifs satellite are shown in (b); the data sets for both (a) and (b) are from Wood *et al.* (2008); the Swiss roll in (c) is a spiralling band in a three-dimensional Euclidean space; the Bitten torus in (d) is constructed by removing the lower right part of a torus; synthetic data sets are considered on the surface of (c) and (d); details for constructing (c) and (d) are available in the on-line supplementary material

the intrinsic geometry of the sea and its complex boundary. Traditional smoothing or modelling methods that do not respect the intrinsic geometry of the space, and in particular the boundary constraints, may produce poor results.

For example, it is crucial to take into account the fact that pairs of locations having close Euclidean distance may be intrinsically far apart if separated by a land barrier. Refer in particular to the locations near longitude 58.5° and 59° in the southern region of the map in Fig. 1(b). These locations have quite different levels of chlorophyll because of the land barrier. However, usual smoothing or modelling approaches that do not account for the boundary would naturally provide close estimates of the level of chlorophyll given their closeness spatially. The goal of this paper is to provide a general methodology that can accommodate not just complex spatial

subregions of $\mathbb{R}^2$ (refer also to the U-shaped constraint in Fig. 1(a)) but also complex subregions of higher dimensional space ($\mathbb{R}^3$ and beyond) and constraints, such as the Swiss roll in Fig. 1(c) and the Bitten torus in Fig. 1(d).

To accommodate modelling on these broad and complex domains, we propose a novel class of *intrinsic* Gaussian processes (GPs). An intrinsic GP refers to a GP that employs the intrinsic Riemannian geometry of the manifold, including the boundary features and interior conditions. Note that this intrinsic notion of intrinsic GPs is different from the *intrinsic random functions* that were defined in the seminal work of Matheron (1973), which refer to processes that have a more general form of stationarity than the usual second-order stationarity. Intrinsic GPs are designed to be useful in interpolation, regression and classification on manifolds, with a particular emphasis on complex or difficult regions arising as submanifolds. A major challenge in constructing GPs on manifolds is choosing a valid covariance kernel—this is a non-trivial problem and most of the focus has been on developing covariance kernels that are specific to a particular manifold (for example, Guinness and Fuentes (2016) considered low dimensional spheres). Castíllo *et al.* (2014) instead proposed to use randomly rescaled solutions of the heat equation to define a valid covariance kernel for reasonably broad classes of compact manifolds. They additionally provided lower and upper bounds on contraction rates of the resulting posterior measure. Unfortunately, they did not provide a methodology for implementing their approach in practice, and their proposed heat kernels are computationally intractable.

This paper proposes a practical and general intrinsic GP methodology, which uses heat kernels as covariance kernels. This is made possible by the major novel contribution of the paper, which is to utilize connections between heat kernels and transition densities of Brownian motion (BM) on manifolds to obtain algorithms for approximating covariance kernels. Specifically, the covariance kernels are approximated by first simulating a BM on the manifold or complex constrained space of interest, and then evaluating the transition density of the BM. The heat kernel generalizes the popular and well-studied squared exponential kernel to the manifold and arises from the Laplace operator, thus fully exploiting the intrinsic geometry of the space. We utilize a discretized version of BM on manifolds (without boundary) or reflective Brownian motion (RBM) for a Riemannian manifold with boundary. RBMs have been defined and thoroughly studied for Euclidean domains (Lions and Sznitman, 1984; Burdzy *et al.*, 2004; Zhou *et al.*, 2017). A C^2-boundary guarantees the existence and uniqueness of an RBM (remark 3 of Zhou *et al.* (2017)). The transition density functions are the Neumann heat kernels of the domain (Hsu, 1984).

Most current methods that can smooth noisy data over regions with a boundary can be applied only to spaces that are *subsets of* $\mathbb{R}^2$; refer to Wood *et al.* (2008) and Ramsay (2002). Sangalli *et al.* (2013) extended Ramsay's (2002) smoothing spline method to model the brain surface arising as a subset of $\mathbb{R}^3$ by first discretizing the surface. The main idea in this literature is to develop smoothing splines that respect the boundary or interior constraints. Our intrinsic GP approach is fundamentally different conceptually, while also having general applicability beyond two-dimensional examples. Although intrinsic GPs have an increasing computational cost as the dimensionality of the space increases, because of the need to simulate BM, there is no discretization of the space unlike methods proposed in Ramsay (2002) and Sangalli *et al.* (2013).

Related work includes Pelletier (2005) who extended kernel regression to a general Riemannian manifold. Bhattacharya and Dunson (2010) modelled a response and covariate on a manifold jointly by using a Dirichlet process mixture model. The focus of our work in contrast aims to generalize the powerful GP model to manifold-valued data. Although GPs have been extensively used in statistics and machine learning (see for example Rasmussen (2004)), these models cannot be directly generalized to model data on manifolds, such as irregular shape

spaces, because of the difficulty of constructing valid covariance kernels. Lin *et al*. (2018) proposed *extrinsic covariance kernels* on general manifolds by first embedding the manifolds onto a higher dimensional Euclidean space, and constructing a covariance kernel on the images after embedding. However, such embeddings are not always available or easy to obtain for complex spaces.

Aumentado-Armstrong and Siddiqi (2017) adopted a related idea of estimating the heat kernel for a sampled manifold (mesh or point cloud) from BM trajectories. No boundary condition is considered. Their approach can be summarized in three steps.

(a) Construct a local surface for approximating the manifold by using moving least squares.
(b) Simulate BM trajectories that are specific to a local surface by using stochastic differential equations with a local metric tensor. To move across different surfaces or charts, iteratively alternate between BM simulation and project the process onto a local surface.
(c) Estimate the heat kernel from the BM trajectories by using kernel density estimation, expressed as a summation of Gaussian kernels.

Each term is calculated on the basis of the Euclidean distance between the BM sample paths and the target points. This is problematic when the Euclidean distance is small but the geodesic is big, e.g. U-shaped domains or regions of a manifold where the curvature is large. Ozakin and Gray (2009) showed that the kernel density estimation estimator is poor and biased in this context.

In our approach, we estimate the heat kernel by simulating BM sample paths on manifolds with or without a boundary by using a global metric tensor. Our way of constructing the heat kernel estimator is completely different from that of Aumentado-Armstrong and Siddiqi (2017). Instead of using an approximation approach, such as relying on kernel density estimation, we develop a direct approach to estimate the heat kernel based on the definition of the BM transition probability on the manifold.

The paper is organized as follows. Section 2 introduces our construction of covariance kernels on manifolds and explores the connection between the heat kernel on a Riemannian manifold and the transition density of BM on the manifold. This connection is utilized in developing practical algorithms for approximating the heat kernel. Inference under intrinsic GPs using the approximated heat kernel, including an extension to *sparse* intrinsic GPs, is explained in Section 2.4. Properties of the heat kernel estimator are discussed in Section 2.5. Sections 3 and 4 illustrate our intrinsic GP methodology with various simulation and data examples. Section 5 contains a discussion. Computational cost and algorithm complexity of the method are discussed in the on-line supplementary material.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets`

2. Intrinsic Gaussian process on manifolds

2.1. Intrinsic Gaussian processes with heat kernel as the covariance kernel

We propose to construct intrinsic GPs on manifolds and complex constrained spaces by using the heat kernel as the covariance kernel. To be more specific, let M be a d-dimensional complete and orientable Riemannian manifold, ∂M its boundary that is continuous and C^1 almost everywhere, Δ_s the Laplacian–Beltrami operator on M and δ the Dirac delta function. A heat kernel of M is a smooth function $K(x, y, t)$ on $M \times M \times \mathbb{R}^+$ that satisfies the heat equation:

$$\frac{\partial}{\partial t}K_{\text{heat}}(s_0,s,t)=\frac{1}{2}\Delta_s K_{\text{heat}}(s_0,s,t),\qquad \lim_{t\to 0}K_{\text{heat}}(s_0,s,0)=\delta(s_0,s),\qquad s_0,s\in M,$$

where the initial condition holds in a distributional sense (Berline *et al.*, 2003). If ∂M is empty, M admits a unique heat kernel. If ∂M is non-empty, multiple heat kernels exist, but the heat kernel becomes unique when we also impose a suitable condition along ∂M, such as the Neumann boundary condition:

$$\frac{\partial K}{\partial \mathbf{n}}=0\qquad \text{along } \partial M, \tag{1}$$

where $\mathbf{n}$ denotes a normal vector of ∂M.

Alternatively, a heat kernel can be viewed as an operator on $L^2(M)$:

$$f\mapsto \int_M K_{\text{heat}}(x,y,t)f(y)\mathrm{d}y, \tag{2}$$

and as such is equivalent to $\exp(\frac{1}{2}t\Delta)f$, with $\mathrm{d}y$ the infinitesimal Riemannian volume. The heat kernel is symmetric with $K_{\text{heat}}(x,y,t)=K_{\text{heat}}(y,x,t)$ and is a positive semidefinite kernel on M for any fixed t, and thus can serve as a valid covariance kernel for a Gaussian process on M. The *Neumann boundary condition* can be expressed as no heat transfer across the boundary ∂M.

If M is a Euclidean space $\mathbb{R}^d$, the heat kernel has a closed form corresponding to a time varying Gaussian function:

$$K_{\text{heat}}(\mathbf{x}_0,\mathbf{x},t)=\frac{1}{(2\pi t)^{d/2}}\exp\left(-\frac{\|\mathbf{x}_0-\mathbf{x}\|^2}{2t}\right),\qquad \mathbf{x}\in\mathbb{R}^d.$$

In addition, the heat kernel of $\mathbb{R}^d$ can be seen as the scaled version of a radial basis function (RBF) kernel (or the popular squared exponential kernel) under different parameterizations:

$$K_{\text{RBF}}(\mathbf{x}_0,\mathbf{x},l)=\sigma_r^2\exp\left(-\frac{\|\mathbf{x}_0-\mathbf{x}\|^2}{2l^2}\right),\qquad \mathbf{x}\in\mathbb{R}^d.$$

Letting $K^t_{\text{heat}}(x,y)=K_{\text{heat}}(x,y,t)$, our intrinsic GP uses $K^t_{\text{heat}}(x,y)$ as the covariance kernel, where the time parameter t of K_{heat} has the same effect as that of the length scale parameter l of K_{RBF}, controlling the rate of decay of the covariance. By varying the time parameter, one can vary the bumpiness of the realizations of the intrinsic GP.

We use intrinsic GPs to develop non-parametric regression and spatial process models on complex constrained domains M. Let $\mathcal{D}=\{(s_i,y_i),i=1,\ldots,n\}$ be the data, with n the number of observations, $s_i\in M$ the predictor or location value of observation i and y_i a corresponding response variable. We would like to do inferences on how the output y varies with the input s, including predicting y-values at new locations s_* that are not represented in the training data set. Assuming Gaussian noise and a simple measurement structure, we let

$$y_i=f(s_i)+\epsilon_i,\qquad \epsilon_i\sim\mathcal{N}(0,\sigma^2_{\text{noise}}),\quad s_i\in M, \tag{3}$$

where σ^2_{noise} is the variance of the noise. This model can be easily modified to include parametric adjustment for covariates x_i, and to accommodate non-Gaussian measurements (e.g. having exponential family distributions). However, we focus on the simple Gaussian case without covariates for simplicity in exposition.

Under an intrinsic GP prior for the unknown function $f:M\to\Re$, we have

$$p(\mathbf{f}|s_1,s_2,\ldots,s_n)=\mathcal{N}(\mathbf{0},\Sigma), \tag{4}$$

where $\mathbf{f}$ is a vector containing the realizations of $f(\cdot)$ at the sample points $s_1, \ldots, s_n$, $f_i = f(s_i)$, and Σ is the covariance matrix of these realizations induced by the intrinsic GP covariance kernel. In particular, the entries of Σ are obtained by evaluating the covariance kernel at each pair of locations, i.e.

$$\Sigma_{ij} = \sigma_h^2 K_{\text{heat}}^t(s_i, s_j). \tag{5}$$

Following standard practice for GPs, this prior distribution is updated with information in the response data to obtain a posterior distribution. Explicit expressions for the resulting predictive distribution are provided in Section 2.3.

Remark 1. We added an additional hyperparameter σ_h^2 by rescaling the heat kernel for extra flexibility. The parameter σ_h^2 plays a similar role to that of the magnitude parameter of an RBF kernel in the Euclidean space. As mentioned above, the parameter t replaces the length scale parameter in an RBF or squared exponential kernel.

The posterior distribution of f evaluated at locations $\mathbf{S} = (s_1, \ldots, s_n)$ has the form

$$\begin{aligned} f(s)|\mathcal{D} &\sim \text{GP}(m_{\text{post}}, \Sigma_{\text{post}}), \\ m_{\text{post}} &= \Sigma_{s,\mathbf{S}}(\Sigma_{\mathbf{S},\mathbf{S}} + \sigma_{\text{noise}}^2 \mathbf{I})^{-1}\mathbf{y}, \\ \Sigma_{\text{post}} &= \Sigma_{s,s} - \Sigma_{s,\mathbf{S}}(\Sigma_{\mathbf{S},\mathbf{S}} + \sigma_{\text{noise}}^2 \mathbf{I})^{-1}\Sigma_{\mathbf{S},s}, \end{aligned}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$.

One of the key challenges for inference using intrinsic GPs with the construction in this section is that *closed form expressions for K_{heat}^t do not exist for general Riemannian manifolds*. Explicit solutions are available only for very special manifolds, such as Euclidean spaces and spheres. Therefore, for most cases, one cannot explicitly evaluate K_{heat}^t or the corresponding covariance matrices. To overcome this challenge and to bypass the need to solve the heat equation directly, we utilize the fact that heat kernels can be interpreted as *transition densities of BM on M*. Our recipe is to simulate BM on M, to evaluate the transition density of the BM numerically and then to use the evaluation to approximate the kernel $K_{\text{heat}}^t(s_i, s_j)$ for any pair (s_i, s_j). The simulation of BM on Riemannian manifolds is discussed in Section 2.2. We also provide some background on Riemannian geometry and stochastic calculus on manifolds.

2.2. Simulating Brownian motion on manifolds

To estimate the transition density of BM on M, we first need to simulate BM sample paths on M. Let $\phi : \mathbb{R}^d \to M$ be a smooth local parameterization of M around $s_0 \in M$. A demonstration of ϕ is depicted in Fig. 2. Let $\mathbf{x}(t_0) \in R^d$ be such that $\phi\{\mathbf{x}(t_0)\} = s_0$. In this paper, we assume that the local parameterization ϕ is known. Examples of ϕ are given in the on-line supplementary material for the Swiss roll and Bitten torus. If ϕ is unknown, Tosi *et al.* (2014) provided an approach to learn ϕ by doing non-linear dimension reduction using latent variable models.

The Riemannian manifold M is equipped with a metric tensor g. For example, if M is a submanifold of a Euclidean space, the induced metric tensor can be described in local co-ordinates as follows:

$$g_{ij}(\mathbf{x}) = \frac{\partial \phi}{\partial x_i}(\mathbf{x}) \, \frac{\partial \phi}{\partial x_j}(\mathbf{x}). \tag{6}$$

Based on its metric tensor, a Riemannian manifold has an associated Laplace–Beltrami operator Δ_s. In local co-ordinates, Δ_s can be written as

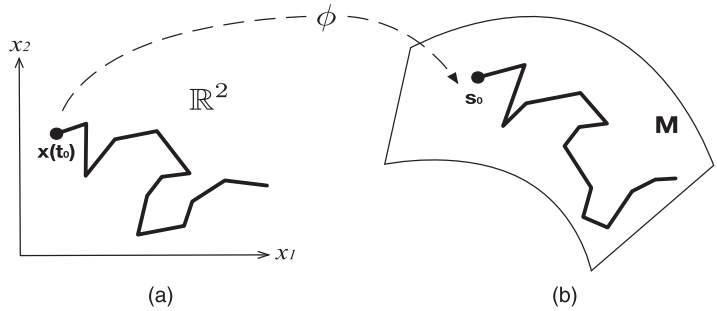

Fig. 2. BM on M and its equivalent stochastic process in a local co-ordinate system in $\mathbb{R}^2$: $\phi : \mathbb{R}^2 \rightarrow M$ is a local parameterization of M

$$\Delta_s f = \frac{1}{\sqrt{G}} \frac{\partial}{\partial x^j} \left(\sqrt{G} g^{ij} \frac{\partial f}{\partial x^i} \right), \tag{7}$$

where G is the determinant of the matrix g and g^{ij} is the (i, j) element of its inverse. As we have seen in Section 2.1, the heat kernel is a solution of the heat equation. The Laplace–Beltrami operator is also the infinitesimal generator of BM on the manifold. Let s_0 be the starting point of the BM S_t on manifold M, and introduce a function $v : M \rightarrow R$. The expectation $E_s\{v(S_t)\} = u(s_0, t)$ satisfies the heat equation, $\partial u / \partial t = \frac{1}{2} \Delta_s u$, $u(s_0, 0) = v(s_0)$.

As in Fig. 2, simulating a sample path of BM on M with starting point s_0 is equivalent to simulating a stochastic process in $\mathbb{R}^d$ with starting point $\mathbf{x}(t_0)$. The BM on a Riemannian manifold in a local co-ordinate system is given as a system of stochastic differential equations in the Ito form (Hsu, 1988, 2008):

$$\mathrm{d}x_i(t) = \frac{1}{2} G^{-1/2} \sum_{j=1}^{d} \frac{\partial}{\partial x_j} (g^{ij} G^{1/2}) \, \mathrm{d}t + (g^{-1/2} \mathrm{d}B(t))_i, \tag{8}$$

where g is the metric tensor of M, G is the determinant of g and $B(t)$ represents independent BM in the Euclidean space. If $M = \mathbb{R}^d$, g becomes an identity matrix and $x_i(t)$ is standard BM in $\mathbb{R}^d$. The first term of equation (8) is related to the local curvature of M. The second term relates to the position-specific alignment of the BM by transforming the standard BM $B(t)$ in $\mathbb{R}^d$ on the basis of the metric tensor g.

For simulating BM sample paths, the discrete form of equation (8) is first derived in equation (9). Specifically, the Euler–Maruyama method is used (Kloeden and Platen, 1992; Lamberton and Lapeyre, 2007), which yields

$$\begin{aligned} x_i(t) &= x_i(t-1) + \frac{1}{2} \sum_{j=1}^{d} \left(-g^{-1} \frac{\partial g}{\partial x_j} g^{-1} \right)_{ij} \Delta t + \frac{1}{4} \sum_{j=1}^{d} (g^{-1})_{ij} \mathrm{tr} \left(g^{-1} \frac{\partial g}{\partial x_j} \right) \Delta t + (g^{-1/2} \mathrm{d}B(t))_i \\ &= \mu(x_i(t-1), \Delta t)_i + (\sqrt{\Delta t} g^{-1/2} z^d)_i, \end{aligned} \tag{9}$$

where Δt is the diffusion time of each step of the BM simulation and z^d represents a d-dimensional standard normal random variable. The discrete form of the above stochastic differential equation defines the proposal mechanism of the BM with density

$$q\{x(t) | x(t-1)\} = \mathbb{N}[x(t) | \mu\{x(t-1), \Delta t\}, \Delta t g^{-1}]. \tag{10}$$

This proposal makes BM move according to the metric tensor. If the manifold M has boundary ∂M, we apply the Neumann boundary condition as in Section 2.1, equation (1). It implies that the simulated sample paths exist only within the boundary.

The simulation of BM from the discretization of stochastic differential equations at points of singularity in the co-ordinate system (e.g. the north pole of a sphere) could be difficult. The drift term (the dt-term in equation (8)) may become too large for the simulated step to be a good approximation of the actual BM. A possible way to address this issue is to limit the size of drift in each simulation step by reducing the time step adaptively.

2.3. Numerical approximation of the heat kernel: exploiting connections with the transition density of Brownian motion

To explain explicitly the equivalence between the heat kernel and the transition density of the BM, let $S(t)$ denote a BM on M started from s_0 at time $t=0$. The probability of $S(t) \in A \subset M$, for any Borel set A, is given by

$$\mathbb{P}\{S(t) \in A | S(0) = s_0\} = \int_A K^t_{\text{heat}}(s_0, s)\mathrm{d}s, \tag{11}$$

where the integral is defined with respect to the volume form of M. In this context, the Neumann boundary condition on the heat kernel corresponds to BM reflecting at the boundary. This can be approximated by pausing time and resampling the next step until it stays within the boundary. The difference between reflecting and resampling is small when the proposed BM step is not far from the boundary. Further discussions are provided in the on-line supplementary material.

We approximate the heat kernel via approximating the integral in equation (11) by simulating BM sample paths and numerically evaluating the transition probability. Considering the BM $\{S(t) : t > 0\}$ on M with the starting point $S(0) = s_0$, we simulate N sample paths. For any $t > 0$ and $s \in M$, the probability of $S(t)$ in a small neighbourhood A of s can be estimated by counting how many BM sample paths reach A at time t. Note that the BM diffusion time t works as the smoothing parameter. If t is large, the BM has higher probability of reaching the neighbourhood of the target point and leads to higher covariance and vice versa. The *transition probability* is approximated as

$$\mathbb{P}\{S(t) \in A \mid S(0) = s_0\} \approx \frac{k}{N}, \tag{12}$$

where N is the number of simulated BM sample paths and k is the number of BM sample paths which reach A at time t. An illustrative diagram is shown in Fig. 3. The *transition density* of $S(t)$ at s is approximated as

$$K^t_{\text{heat}}(s_0, s) \approx \hat{K}^t = \mathbb{P}\{S(t) \in A \mid S(0) = s_0\} \approx \frac{1}{V(A)}\frac{k}{N}, \tag{13}$$

where $V(A)$ is the Riemannian volume of A, which is parameterized with the radius of A, and $\hat{K}^t$ is the estimated transition density. The error (numerical and Monte Carlo) of this estimator of the heat kernel is discussed in Section 2.5.

Remark 2. We are not aware of any rigorous definition of RBM for a general Riemannian manifold *with boundary*. We conjecture that, given a suitable definition of RBM in a general Riemannian manifold with boundary (that generalizes the existing definition for a Euclidean domain), the RBM exists and is unique if the boundary is C^2 (almost everywhere), and its transition density functions are the Neumann heat kernels.

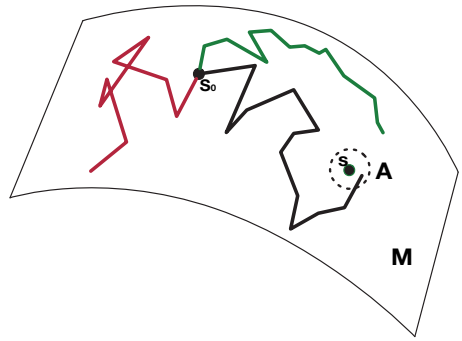

Fig. 3. BM on a manifold M: s_0 is the starting point of BM sample paths (in this example, only the black sample path reaches A at time t and the estimate of the transition probability $p\{S(t) \in A | S(0) = s_0\}$ is $\frac{1}{3}$: ———, ———, ———, three independent BM sample paths from time 0 to t; ◌, a set A, which is a neighbourhood of a point s on M

The intrinsic GP can be constructed by using approximation (13). The covariance matrix of the training data $\Sigma_{\mathbf{ff}}$ can be explicitly obtained as follows: for the ith row of $\Sigma_{\mathbf{ff}}$, N BM sample paths are simulated, with the starting point the ith data point indexed by the corresponding row. For each element of the ith row, $\Sigma_{\mathbf{ff}}$, $\hat{K}^{\mathrm{t}}(s_i, s_j)$ is then estimated by using expression (13). Algorithm 1 in Table 1 provides details on how to generate $\Sigma_{\mathbf{ff}}$.

Optimization of the kernel hyperparameters is discussed in Section 2.6. Given intrinsic GPs as the prior, we can then update with the likelihood to obtain the posterior distribution for inference. Let $\mathbf{f}_*$ be a vector of values of $f(\cdot)$ at some test points that are not represented in the training sample. The joint distribution of $\mathbf{f}$ and $\mathbf{f}_*$ is

$$p(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left\{0, \begin{pmatrix} \Sigma_{\mathbf{ff}} & \Sigma_{\mathbf{ff}_*} \\ \Sigma_{\mathbf{f}_*\mathbf{f}} & \Sigma_{\mathbf{f}_*\mathbf{f}_*} \end{pmatrix}\right\}, \tag{14}$$

where $\Sigma_{\mathbf{f}_*\mathbf{f}}$ is the covariance matrix for training and test data points. Each entry of the covariance matrix of the joint distribution can be calculated by using equation (15):

$$\Sigma_{ij} = \sigma_h^2 \hat{K}^{\mathrm{t}}(s_i, s_j). \tag{15}$$

Table 1. Algorithm 1: simulating BM sample paths for estimating Σ

Step 1(a): generate BM sample paths
for $i = 1, \ldots, N_{\mathrm{d}}$ *do* (N_{d} is the size of data points)
  for $j = 1, \ldots, N_{\mathrm{BM}}$ *do* (N_{BM} is the number of sample paths)
    for $l = 1, \ldots, T$ *do* (T-steps BM, $T\Delta t \to$ maximum diffusion time)
      do (keep proposing x until the value is within the boundary)
      $q\{x_{i,j}(l) | x_{i,j}(l-1)\} \leftarrow \mathbb{N}[x_{i,j}(l) | \mu\{x_{i,j}(l-1), \Delta t\}, \Delta t g^{-1}]$ (use equation (10))
      while $x_{i,j}$ is located outside M
return x
Step 1(b): given a discrete choice of the diffusion time $t \in \{\Delta t, 2\Delta t, \ldots, T\Delta t\}$, the covariance matrix Σ^t is estimated on the basis of the BM simulation from step 1(a)
for $i = 1, \ldots, N_{\mathrm{d}}$ *do*
  for $j = 1, \ldots, N_{\mathrm{d}}$ *do*
    $k =$ which $\{x(t) \in A_j\}$ (counting how many BM paths reach A_j)
    $K^{\mathrm{t}}_{\mathrm{heat}}(s_i, s_j) = k/\{N_{\mathrm{BM}} V(A_j)\}$ (use equation (13))
    $\Sigma^{\mathrm{t}}_{ij} = \sigma_h^2 K^{\mathrm{t}}_{\mathrm{heat}}(s_i, s_j)$
return Σ^{t}

For the same row of $\Sigma_{\mathbf{ff}}$ and $\Sigma_{\mathbf{ff}_*}$, all elements can be estimated from the same patch of BM simulations which share the same starting points. No additional BM simulations are needed to estimate $\Sigma_{\mathbf{ff}_*}$. The predictive distribution is derived by marginalizing out $\mathbf{f}$:

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*\mathbf{f}|\mathbf{y})\mathrm{d}\mathbf{f} = \mathcal{N}\{\Sigma_{\mathbf{f}_*\mathbf{f}}(\Sigma_{\mathbf{ff}} + \sigma^2_{\text{noise}}I)^{-1}\mathbf{y}, \Sigma_{\mathbf{f}_*\mathbf{f}_*} - (\Sigma_{\mathbf{ff}} + \sigma^2_{\text{noise}}I)^{-1}\Sigma_{\mathbf{ff}_*}\}.$$

If we are interested only in the predictive mean, only $\Sigma_{\mathbf{f}_*\mathbf{f}}$ and $\Sigma_{\mathbf{ff}}$ need to be estimated. The predictive variance of test points requires computing the covariance matrix $\Sigma^t_{\mathbf{f}_*\mathbf{f}_*}$. This requires extra BM simulations whose starting points are the test points. This could be computationally heavy if the number of test points is big. The sparse intrinsic GP is introduced in the next section to handle this problem.

2.4. Sparse intrinsic Gaussian process on manifolds to reduce computation cost

The construction of intrinsic GPs proposed in Section 2.3 requires simulating BM sample paths at each data point. Although the BM simulations are embarrassingly (or trivially parallelizable, the computational cost can be high when the sample size is large. In addition, GPs face the well-known problem of high computational complexity $O(n^3)$ due to the inversion of the covariance matrix. In this section, we propose to combine intrinsic GPs with sparse GP approximations proposed by Quiñonero-Candela *et al.* (2007). We call the resulting construction a *sparse intrinsic GP*. By employing a sparse intrinsic GP, BM paths only need to be simulated starting at the induced points instead of every data point. The intuition behind this is that many training data are located close together, implying that there may be a large amount of redundant information. The inducing point approximation summarizes the training data into a small set of inducing points, so that inference could be done more efficiently.

The GP prior can be augmented with an additional set of m inducing points on M denoted as $\mathbf{z} = (z_1, \ldots, z_m)$, $z_i \in M$, and we have m random variables $\mathbf{u} = (f(z_1), \ldots, f(z_m))$. The marginal prior distribution $p(\mathbf{f}_*, \mathbf{f})$ remains unchanged after the model has been rewritten in terms of the prior distribution $p(\mathbf{u})$ and the conditional distribution $p(\mathbf{f}_*, \mathbf{f}|\mathbf{u})$:

$$p(\mathbf{f}_*, \mathbf{f}) = \int p(\mathbf{f}_*, \mathbf{f}, \mathbf{u})\mathrm{d}\mathbf{u} = \int p(\mathbf{f}_*, \mathbf{f}|\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{u}, \qquad p(\mathbf{u}) = \mathcal{N}(0, \Sigma_{\mathbf{uu}}), \tag{16}$$

where the distribution of u is multivariate Gaussian with mean 0 and covariance matrix $\Sigma_{\mathbf{uu}}$. The above augmentation does not reduce the computational complexity. For efficient inference, we adopt the deterministic inducing conditional approximation by Quiñonero-Candela *et al.* (2007), where $\mathbf{f}_*$ and $\mathbf{f}$ are assumed to be conditionally independent given $\mathbf{u}$ and the relationships between any $\mathbf{f}$ and $\mathbf{u}$ are deterministic:

$$p(\mathbf{f}_*, \mathbf{f}) \approx q(\mathbf{f}_*, \mathbf{f}) = \int q(\mathbf{f}_*|\mathbf{u})q(\mathbf{f}|\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{u}, \tag{17}$$

$$q(\mathbf{f}|\mathbf{u}) = \mathbf{N}(\boldsymbol{\mu}_\mathbf{f}, 0), \qquad \boldsymbol{\mu}_\mathbf{f} = \Sigma_{\mathbf{fu}}\Sigma^{-1}_{\mathbf{uu}}\mathbf{u}, \tag{18}$$

$$q(\mathbf{f}_*|\mathbf{u}) = \mathbf{N}(\boldsymbol{\mu}_*, 0), \qquad \boldsymbol{\mu}_* = \Sigma_{\mathbf{f}_*\mathbf{u}}\Sigma^{-1}_{\mathbf{uu}}\mathbf{u}. \tag{19}$$

The resulting sparse intrinsic GP prior is

$$q(\mathbf{f}, \mathbf{f}_*) = \mathbf{N}\left\{0, \begin{pmatrix} Q_{\mathbf{ff}} & Q_{\mathbf{ff}_*} \\ Q_{\mathbf{f}_*\mathbf{f}} & Q_{\mathbf{f}_*\mathbf{f}_*} \end{pmatrix}\right\} = \mathbf{N}\left\{0, \begin{pmatrix} \Sigma_{\mathbf{fu}}\Sigma^{-1}_{\mathbf{uu}}\Sigma_{\mathbf{uf}} & \Sigma_{\mathbf{fu}}\Sigma^{-1}_{\mathbf{uu}}\Sigma_{\mathbf{uf}_*} \\ \Sigma_{\mathbf{f}_*\mathbf{u}}\Sigma^{-1}_{\mathbf{uu}}\Sigma_{\mathbf{uf}} & \Sigma_{\mathbf{f}_*\mathbf{u}}\Sigma^{-1}_{\mathbf{uu}}\Sigma_{\mathbf{uf}_*} \end{pmatrix}\right\},$$

where Q is defined as $Q_{\mathbf{a},\mathbf{b}} = \Sigma_{\mathbf{a},\mathbf{u}}\Sigma_{\mathbf{u},\mathbf{u}}^{-1}\Sigma_{\mathbf{u},\mathbf{b}}$. Using algorithm 1, $\Sigma_{\mathbf{uu}}$, $\Sigma_{\mathbf{uf}}$ and $\Sigma_{\mathbf{uf}_*}$ are all obtained by estimating the transition density of BM simulation paths with inducing points as the starting points.

We then only need to simulate the BM sample paths starting from the inducing points. The total number of BM simulations is reduced from $n \times N_{\mathrm{BM}}$ to $m \times N_{\mathrm{BM}}$, where m is the number of inducing points, n is the number of data points and N_{BM} is the number of BM sample paths given a single starting point. The complexity of inverting the covariance matrix is decreased from $O(n^3)$ to $O(n \times m^2)$.

With the above approximation, the marginal distribution of the corresponding GP with a Gaussian likelihood is written as

$$p(\mathbf{y}|\mathbf{f}) \approx q(\mathbf{y}|\mathbf{u}) = \prod_{i=1}^{n} \mathbf{N}(y_i|\Sigma_{f_i\mathbf{u}}\Sigma_{\mathbf{uu}}^{-1}\mathbf{u}, \sigma_{\mathrm{noise}}^2\mathbf{I}). \tag{20}$$

The inducing points in the above marginal likelihood can be further marginalized out by substituting the definition of its prior distribution (16):

$$p(\mathbf{y}|\mathbf{s}_{\mathrm{induce}}) = \int q(\mathbf{y}|\mathbf{u})\, p(\mathbf{u}|\mathbf{s}_{\mathrm{induce}})\mathrm{d}\mathbf{u} = \mathbf{N}(0, \Sigma_{\mathbf{fu}}\Sigma_{\mathbf{uu}}^{-1}\Sigma_{\mathbf{uf}} + \sigma_{\mathrm{noise}}^2\mathbf{I}). \tag{21}$$

With this model, we can also obtain the predictive distribution as

$$q(\mathbf{f}_*|\mathbf{y}) = \mathbf{N}\{Q_{\mathbf{f}_*\mathbf{f}}(Q_{\mathbf{ff}} + \sigma^2\mathbf{I})^{-1}\mathbf{y},\, Q_{\mathbf{f}_*\mathbf{f}_*} - Q_{\mathbf{f}_*\mathbf{f}}(Q_{\mathbf{ff}} + \sigma^2\mathbf{I})^{-1}Q_{\mathbf{ff}_*}\}. \tag{22}$$

There is a huge literature on reducing the matrix inversion bottleneck in GP computation (Schwaighofer and Tresp, 2002; Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006; Titsias, 2009). Recent approaches, such as Katzfuss and Guinness (2017), can achieve linear time computation complexity under certain conditions. However, such approaches require an analytical form of covariance kernel; to apply these methods, we would need to simulate BM paths at the training and prediction points. For this reason, we use the deterministic inducing conditional because of its avoidance of the need to estimate the diagonal elements of the covariance matrix.

2.5. Monte Carlo and numerical error for approximation of heat kernel

In this section, we discuss the error of our heat kernel estimator as defined in equation (13). We also consider numerical experiments in the special case of $\mathbb{R}$ in which case the true heat kernel is known.

Consider BM $\{S(t) : t > 0\}$ on a Riemannian manifold M with $S(0) = s_0$. Fix some $t > 0$ and $s \in M$. The probability density of $S(t)$ at s is $K_{\mathrm{heat}}^{\mathrm{t}}(s_0, s)$. The true BM transition probability evaluated at a set A is given by $p(A) = \mathbb{P}\{S(t) \in A | S(0) = s_0\} = \int_A K_{\mathrm{heat}}^{\mathrm{t}}(s_0, s)\mathrm{d}s$. The error of our estimator $\hat{K}^{\mathrm{t}}$ consists of two parts.

2.5.1. Part I: numerical error

Choose local co-ordinates $(r_1, \ldots, r_d)$ near s with $r_1(s) = \ldots = r_d(s) = 0$ (for convenience of illustration) and a window size w. The heat kernel $K_{\mathrm{heat}}^{\mathrm{t}}$ can then be approximated by

$$K^{\mathrm{t}\prime} = \frac{1}{V(A)}\mathbb{P}[|r_i\{S(t)\}| < w \text{ for } i = 1, \ldots, d],$$

where $V(A)$ denotes the volume of the region defined by $\{|r_i| < w, i = 1, \ldots, d\}$. By Taylor series expansion around s, we have

$$K^{t'} = K^{t}_{\text{heat}} + O(w^2). \tag{23}$$

Therefore, the approximation error increases (quadratically) with w, i.e. the order of magnitude of $K^{t'} - K^{t}_{\text{heat}}$ is $O(w^2)$.

If $M = R^d$, we can explicitly derive the error. Assuming that the starting point of BM s_0 is the origin for simplicity, the heat kernel K^{t}_{heat} on R^d can be approximated as

$$\begin{aligned} K^{t'} &= \frac{1}{V(A)} \mathbb{P}\{\|S(t) - s\| < w\} = \frac{1}{V(A)} \int_A K^{t}_{\text{heat}}(s_0, s) \mathrm{d}s, \\ &= \frac{1}{(2w)^d} \int_{s_1 - w}^{s_1 + w} \cdots \int_{s_d - w}^{s_d + w} \exp\left(\frac{-\sum_{i=1}^{d} x_i^2}{2t} \right) \mathrm{d}x_d \ldots \mathrm{d}x_1. \end{aligned} \tag{24}$$

Taylor series expansion of equation (24) yields

$$K^{t'} - K^{t}_{\text{heat}} = \frac{\sum_{i=1}^{d} s_i^2 - \mathrm{d}t}{6t} \frac{w^2}{t} + O\left(\frac{w^4}{t^2} \right). \tag{25}$$

Assuming that w is small compared with $\sqrt{t}$, the order of magnitude of this error is $O(w^2)$.

Remark 3. For convenience in computing the integral in equation (24), a hypercube is used instead of the Euclidean ball. The order of magnitude of the error remains the same.

2.5.2. *Part II: Monte Carlo error*

Given N_{BM}, the number of BM sample paths, $K^{t'}$ is approximated by $\hat{K}^{t}$:

$$\hat{K}^{t} = \frac{1}{V(A)} \frac{k}{N_{\text{BM}}}, \qquad k \sim \text{Bin}\{N_{\text{BM}}, V(A) K^{t'}\}. \tag{26}$$

Recall that k is the number of sample paths within $\|S(t) - s\| < w$ and has binomial distribution with N_{BM} trials and probability of success $V(A) K^{t'}$. Here k/N_{BM} is the estimate of the transition probability of BM.

The expectation and the standard error of $\hat{K}^{t}$ are

$$E\{\hat{K}^{t}(s_0, s)\} = K^{t'} = K^{t}_{\text{heat}}(s_0, s) + O(w^2), \tag{27}$$

$$\begin{aligned} \text{sd}(\hat{K}^{t}) &= \frac{1}{N_{\text{BM}} V(A)} \sqrt{[N_{\text{BM}}\, V(A) K^{t'} \{1 - V(A) K^{t'}\}]} \\ &\leqslant \sqrt{\left\{ \frac{K^{t'}}{N_{\text{BM}} V(A)} \right\}} = O(N_{\text{BM}}^{-1/2} w^{-d/2}), \end{aligned} \tag{28}$$

The standard deviation decreases with w and N_{BM}. As $w^2 \to 0$, $E\{\hat{K}^{t}(s_0, s)\} = K^{t}_{\text{heat}}(s_0, s)$, and also as $w^{-d}/N_{\text{BM}} \to 0$, $\text{var}\{\hat{K}^{t}(s_0, s)\} = 0$. The estimator $\hat{K}^{t}(s_0, s)$ is asymptotically unbiased and consistent.

The optimal order of magnitude of w_{opt} can be calculated by minimizing the sum of the numerical error and Monte Carlo error as described above. Specifically, for an arbitrary M, given a fixed number of BM simulations N_{BM}, we have

$$\mathcal{L}(w) = O(w^2) + O(w^{-d/2}). \tag{29}$$

In particular if M is R^d, an explicit expression of the error is available:

$$\mathcal{L}(w) = \sqrt{\left\{ \frac{\hat{K}^{\mathrm{t}}}{N(2w)^d} \right\} + K^{\mathrm{t}}_{\mathrm{heat}} \frac{\sum_{i=1}^{d} s_i^2 - \mathrm{d}t}{6t} \frac{w^2}{t}}. \tag{30}$$

Given a prespecified error level, the order of the minimum number of BM simulations N required can be derived. Refer to the on-line supplementary material for the example of estimating the heat kernel in a one-dimensional Euclidean space.

Numerical accuracy of estimates for the special case of $\mathbb{R}$ are shown in Table 2 and Fig. 4. The true heat kernel $K^{\mathrm{t}}_{\mathrm{heat}}(0, s)$ is calculated by using equation (1) in the on-line supplementary material at 70 equally spaced $s \in (-9, 9)$. The diffusion time is fixed as 10. The transition probability of BM from the origin to the grid point s is estimated by counting how many BM paths reach the neighbourhood of s ($[s-w, s+w]$) at time t. The transition density of BM at each grid

Table 2. Comparison of estimates of BM transition density and the heat kernel in $\mathbb{R}$†

Number of sample paths N_{BM}	*Median absolute error*	*Median relative error (%)*
3×10^2	8.4×10^{-3} (8.9×10^{-3})	24.6 (25.6×10^{-1})
3×10^3	2.8×10^{-3} (2.9×10^{-3})	6.4 (5.5×10^{-2})
3×10^4	7.2×10^{-4} (6.8×10^{-4})	1.6 (1.9×10^{-2})
3×10^5	4.7×10^{-4} (3.8×10^{-4})	1.3 (1.1×10^{-2})

†The table shows the median absolute error and median relative error between the true heat kernel $K^{\mathrm{t}}_{\mathrm{heat}}$ and the numerical estimate of the BM transition density. Values in parentheses show the median absolute deviation.

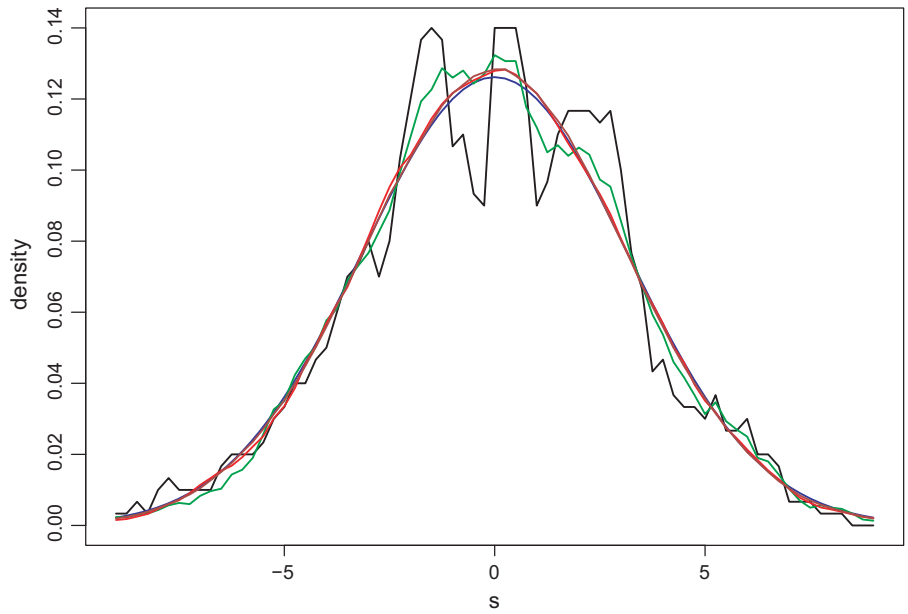

Fig. 4. Comparison of estimates of BM transition density and the heat kernel in $\mathbb{R}$: ——, true heat kernel; ——, 300 BM simulations; ——, 3000 BM simulations; ——, 30000 BM simulations; ——, 300000 BM simulations

point is then evaluated by using equation (13). Using equation (7) in the on-line supplementary material the order of magnitude of w_{opt} is derived as 10^{-1}. We fix the radius w as 0.5 in equation (13).

The number of BM simulation sample paths N_{BM} is selected from 300 to 300000 with increasing order of magnitude. The median of relative error decreases as N_{BM} increases and stabilizes after 30000. A similar pattern is observed for the median absolute error. Derivations for the transition density estimate of heat kernel in $\mathbb{R}^2$ are shown in the supplementary material.

2.6. Optimizing the kernel hyperparameters and comparison with a reflective Brownian motion kernel in $\mathbb{R}$

Given a diffusion time t, using algorithm 1 we can generate a covariance matrix $\Sigma_{\mathbf{ff}}^t$ for the training data indexed by t. The log-marginal-likelihood function (over f) is given by (Rasmussen, 2004)

$$p(\mathbf{y}|s) = \int p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f}|s)\mathrm{d}\mathbf{f} = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}(\Sigma_{\mathbf{ff}}^t + \sigma_{\text{noise}}^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|\Sigma_{\mathbf{ff}}^t + \sigma_{\text{noise}}^2 I| - \frac{N_d}{2}\log(2\pi). \quad (31)$$

The hyperparameters can be obtained by maximizing the logarithm of the marginal likelihood. The maximum of the BM diffusion time is set as $T\,\Delta t$, where T is a positive integer, and Δt is the BM simulation time step as defined in expression (10). T covariance matrices $\Sigma_{\mathbf{ff}}^{1..T}$ can be generated on the basis of the BM simulations. Optimization of diffusion time t can be done by selecting the corresponding $\Sigma_{\mathbf{ff}}^t$ that maximizes the log-marginal-likelihood. Estimation of σ_h given the smoothing parameter t follows by using standard optimization routines, such as quasi-Newton optimization. For the sparse intrinsic GP, the likelihood function is replaced by equation (21) and the hyperparameters can be obtained by similar procedures.

We compare the estimates of kernel hyperparameters from a Euclidean GP (the standard GP in the Euclidean space) and the intrinsic GP in $\mathbb{R}$ by applying both methods to 10 sets of test data. Data sets are generated by sampling 20 data points from a multivariate normal distribution with mean 0 and covariance Σ_{test}. Σ_{test} is produced by a standard RBF kernel with $l = 1$ and $\sigma_r = 1$. In this case, the ground truth of the hyperparameters of the heat kernel is known.

We simulate $N_{\text{BM}} = 40000$ BM sample paths for each test data point. The estimates of hyperparameters t and σ_h are obtained by maximizing equation (31). For the case of $\mathbb{R}$, the two methods should produce very similar results, since the heat kernel is equivalent to an RBF kernel in $\mathbb{R}$.

The result is shown in Table 3, which records the true value and the median estimates of kernel hyperparameter l and σ. Values in parentheses show the median absolute deviation. The error bounds provide insights about the level of error that is introduced by a random walk. The p-values of Wilcoxon tests indicate that the difference in medians between the two methods are not significant.

Table 3. Comparison of estimates of kernel hyperparameters from the Euclidean GP and intrinsic GP in $\mathbb{R}$

Case	*Median estimates of* l	*Median estimates of* σ_r
Truth	1	1
Euclidean GP	1.13 (0.16)	0.94 (0.36)
Intrinsic GP	1.15 (0.2)	0.94 (0.38)
p-value	0.91	0.85

3. Simulation studies

In this section, we carry out simulation studies for a regression model with true regression functions defined on a U-shaped domain, a two-dimensional Swiss roll embedded in $\mathbb{R}^3$ and the Bitten torus. The performance of the intrinsic GP is compared with that of a Euclidean GP (the standard GP as in Rasmussen (2004)) and the soap film smoother in Wood *et al*. (2008) for the U-shape example. For the Swiss roll and Bitten torus examples, the results from an intrinsic GP are compared with those from a Euclidean GP model. Examples of BM sample paths on the U-shape domain, Swiss roll and Bitten torus are shown in Fig. 5.

3.1. U-shape example

A U-shaped domain (see for example Wood (2001)), defined as a subset of $\mathbb{R}^2$, is plotted in

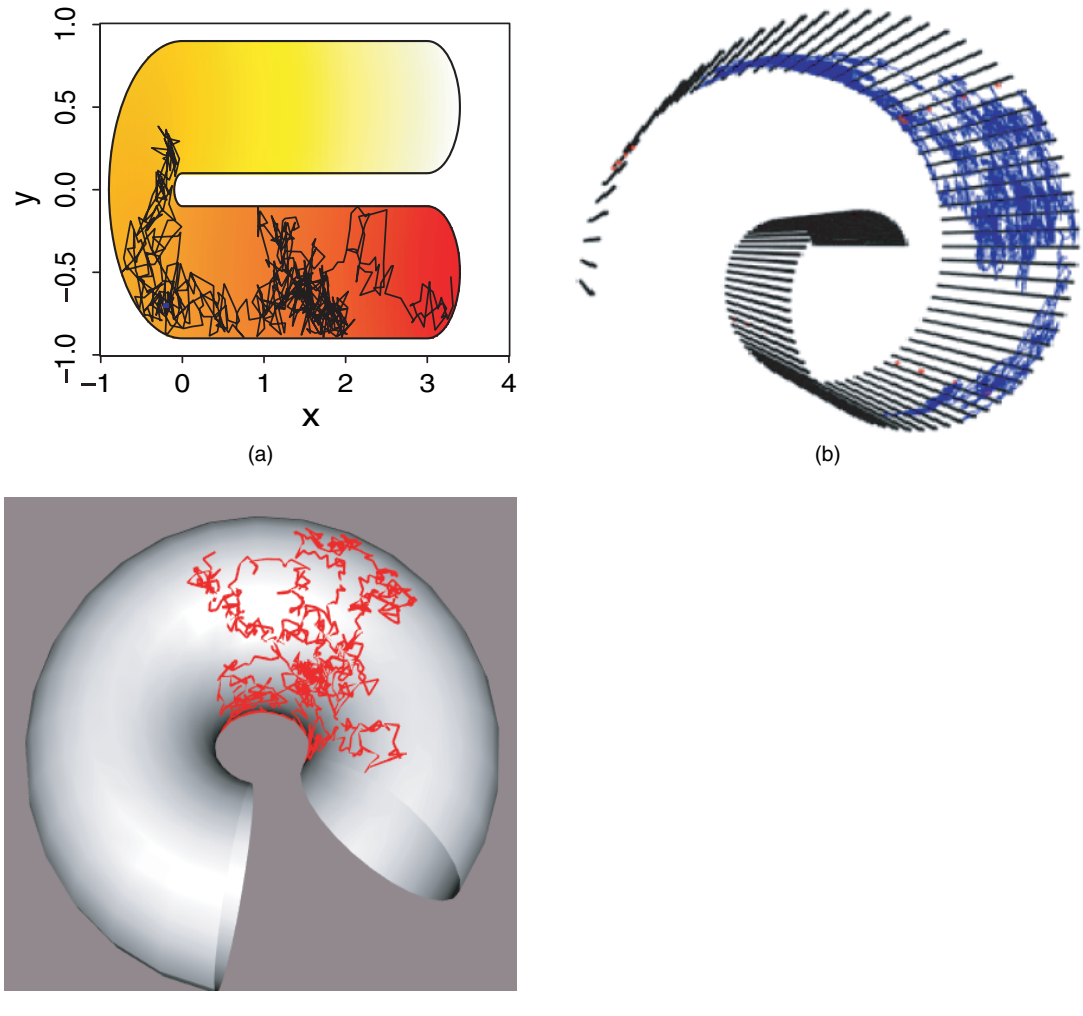

Fig. 5. Examples of BM sample paths on (a) the U-shape domain (———), on (b) the Swiss roll (———) and on (c) the Bitten torus (———)

Table 4. Comparison of the root-mean-squared error of predictive means for various methods on the U-shape domain†

Case	*Result for the following methods:*		
	Euclidean GP	*Intrinsic GP*	*Soap film smoother*
30 db	1 (0.01)	0.274 (0.04)	0.271 (0.22)
10 db	1.36 (0.17)	0.754 (0.14)	0.747 (0.37)

†The table shows the mean of the root-mean-squared error over 50 data sets. Values in parentheses show the standard deviation.

Fig. 6(a). The value of a test or regression function (i.e. the colour of the map) varies smoothly from the lower right-hand corner to the upper right-hand corner of the domain ranging from -6 to 6. The black crosses represent 20 observations which were equally spaced in both x- and y-directions within the domain of interest. The goal is to estimate the test function and to make predictions at 450 equally spaced grid points within the domain.

Since the U-shaped domain is defined as a subset of $\mathbb{R}^2$, the mapping function ϕ in equation (2) is a constant. Therefore, BM reduces to standard BM in the two-dimensional Euclidean space restricted within the boundary. When a proposed BM step hits the boundary, the proposed move is rejected. New proposal steps will be made until the proposed sample path locates within the boundary. The trajectory of a sample path (the black line) of the BM is shown in Fig. 5(a) with the blue dot serving as the starting location.

The heat map of the predictive mean of an intrinsic GP at the grid points is shown in Fig. 6(e). The coloured contours of the prediction are similar to that of the true function in Fig. 6(a). The contours of the Euclidean GP predictive mean in Fig. 6(c) are more squashed, and the differences are exacerbated when certain observations are removed as in Fig. 6(b). It is clear that the Euclidean GP smooths across the gap between the two arms of the domain (see Fig. 6(d)). This is because the upper arm and lower arm are close in Euclidean distance. In contrast, the intrinsic GP, which takes into account the intrinsic geometry, does not smooth across the gap as seen in Fig. 6(f). Given a fixed diffusion time, the transition probability of BM from points in the lower arm to points in the upper arm within the boundary is relatively small. This leads to lower covariance between these two regions and more accurate predictions.

The U-shaped domain example has also been used for evaluating the performance of the soap film smoothers in Wood *et al*. (2008), in comparison with some other methods such as thin plate splines and the method of Ramsay (2002). Comparisons that were made in Wood *et al*. (2008) show that the soap film smoother outperforms the other methods. In our study, the intrinsic GP, Euclidean GP and soap film smoother are compared for various levels of signal-to-noise ratio. The values of the true function are perturbed by Gaussian noise with a standard deviation of 0.1 and 1 (the signal-to-noise ratios are 30 db and 10 db respectively) with 50 replicates for each level of noise. For each of the replicates, different methods are applied to estimate the test function at the grid points. The mean and standard deviation of the mean-squared error for these 50 replicates are reported in Table 4. The soap film smoother is constructed by using 10 inner knots and 10 cubic splines. The intrinsic GP and soap film are both significantly better than the Euclidean GP. There is no substantial difference in terms of the mean mean-squared error between the two methods. However, the standard deviation of the mean-squared error for the intrinsic GP is substantially smaller for both levels of noise. This indicates that the prediction of the intrinsic GP is more robust.

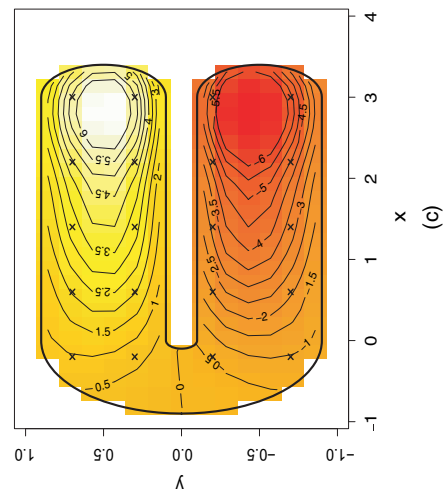

Fig. 6. Comparison of the intrinsic GP and the Euclidean GP in the U-shaped example: (a) true function and data points (the contours are more squashed than in (e)); (b) true function with fewer data points removed from the upper right arm; (c) GP prediction with all data points; (d) Euclidean GP prediction with fewer data points; (e) intrinsic GP prediction with all data points; (f) intrinsic GP prediction with fewer data points (the differences are exacerbated; it is clear that the Euclidean GP smooths across the boundary in (d))

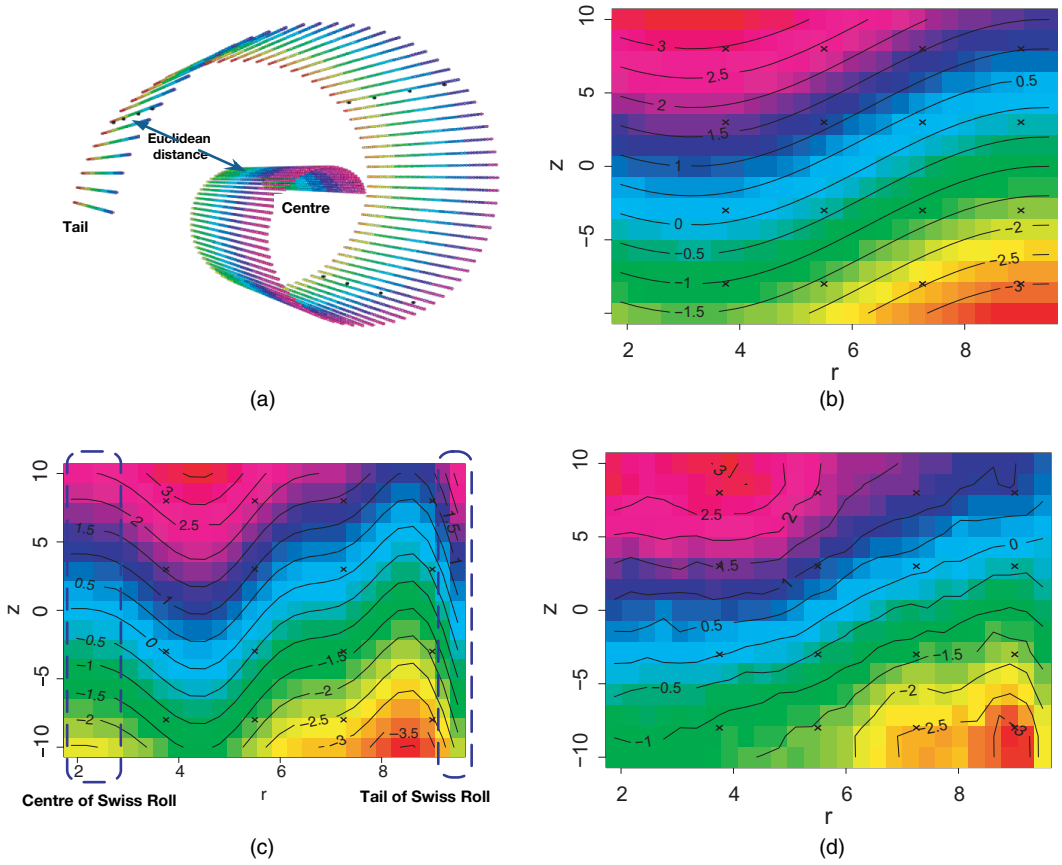

Fig. 7. Comparison of the Euclidean GP and intrinsic GP on a Swiss roll: (a) true function and data points (●) plotted on the surface of a Swiss roll; (b) true function plotted in the two-dimensional unfolded Swiss roll with co-ordinates r and z; (c) GP predictions plotted with contours (the left-hand end and right-hand end marked by a blue broken box are quite different from (b) in colour); (d) instrinsic GP prediction (the prediction at the centre and the tail part of the Swiss roll (the right-hand end and left-hand end) has been improved)

3.2. Swiss roll

The intrinsic GP model applies to general Riemannian manifolds and has much wider applicability to complex spaces beyond subsets of $\mathbb{R}^2$. Here we consider a synthetic data set on a Swiss roll which is a two-dimensional manifold embedded in $\mathbb{R}^3$. The soap film method is only appropriate for smoothing over regions of $\mathbb{R}^2$ and hence cannot be applied here. A Swiss roll is a spiralling band in a three-dimensional Euclidean space. A non-linear function f is defined on the surface of a Swiss roll with

$$Y_i = f(x_i, y_i, z_i) + \epsilon_i,$$

where x_i, y_i and z_i are the co-ordinates of a point on the surface. The construction of the Swiss roll and the derivation of the metric tensor are shown in the on-line supplementary material.

The true function f is plotted in Fig. 7(a). 20 equally spaced observations are marked with black crosses. For better visualization, the true function is plotted in the unfolded Swiss roll in the radius r- and width z-co-ordinates in Fig. 7(b). The true function values are indicated by colour and with contours at the grid points.

We first applied a Euclidean GP to this example by using an RBF kernel in $\mathbb{R}^3$. To visualize the differences between the prediction and the true function, the GP predictive mean is plotted in the unfolded Swiss roll in Fig. 7(c). The overall shape of contours is more wiggly compared with the true function in Fig. 7(b). In addition, the predictive mean is quite different from the truth in colour in certain regions, e.g. the left-hand end of Fig. 7(c) marked by the blue broken box corresponding to the centre of the Swiss roll and the right-hand end of Fig. 7(c) corresponding to the tail of the Swiss roll. The prediction performance of the Euclidean GP in these regions is poor. This is because the Euclidean distance between the two regions is small as shown in Fig. 7(a) whereas the geodesic distance between them (defined on the surface of the Swiss roll) is big.

In applying the intrinsic GP to these data, the BM sample paths can be simulated by using equation (9) using the metric tensor of the Swiss roll. In particular, BM on the Swiss roll can be modelled as the stochastic differential equation

$$\mathrm{d}r(t) = \frac{-2r}{(1+r^2)^2}\mathrm{d}t + \frac{1}{2}\frac{2r}{(1+r^2)^2}\mathrm{d}t + (1+r^2)^{-1/2}\mathrm{d}B_r(t), \tag{32}$$

$$\mathrm{d}z(t) = \mathrm{d}B_z(t), \tag{33}$$

where $B_r(t)$ and $B_z(t)$ are two independent BMs in Euclidean space. A trace plot of a single BM sample path is shown in Fig. 5(b). Following the procedure that was introduced in Section 2.1, the predictive mean of the intrinsic GP is shown in Fig. 7(d). The overall shape of the contour of the predictive mean is similar to that of the true function. The prediction at the centre and tail part of the Swiss roll has been improved compared with the results of the Euclidean GP. The root-mean-square error is calculated between the predictive mean and the true value at the grid points. It has been reduced from 0.53 (Euclidean GP) to 0.29 for the intrinsic GP. The Neumann condition states that at any boundary point the heat kernel is stationary along the normal direction. This directly implies that the level curves of the intrinsic GP prediction are orthogonal to the boundary. However, when the training data are far from the boundary, the level curve can be parallel to the boundary. For example, the right-hand part of Fig. 7(d) corresponds to the tail part of the Swiss roll. When r is big the distance on the surface of the Swiss roll is bigger. The training data are far from the boundary in the tail, so the intrinsic GP prediction tends to be close to the prior mean.

3.3. Bitten torus

Here we consider another more substantial example: a Bitten torus. The torus is a two-dimensional manifold embedded in R^3. The three-dimensional co-ordinates can be parameterized by four variables: r, the radius of the tube, R, the distance from the centre of the tube to the centre of the torus, and (θ, ϕ) angles to parameterize the two full circles with θ for the angle of the torus and ϕ for the angle of the tube. In our case, we fix R and r and vary θ and ϕ. We removed the lower right-hand part to construct the Bitten torus. The Bitten torus is not as 'flat' as the other examples that were considered above.

The value of the test function (i.e. the colour of the map; low values in dark blue and high values in dark red in Fig. 8(a)) increases smoothly from 0.57 to 5.5 on the surface of the Bitten torus. The true function and the noisy observations are plotted in Fig. 8(a). 19 observations are marked with orange balls. 18 of the observations are evenly spaced and one additional observation is near the centre of the Bitten torus. Similarly to the Swiss roll example, the non-linear function

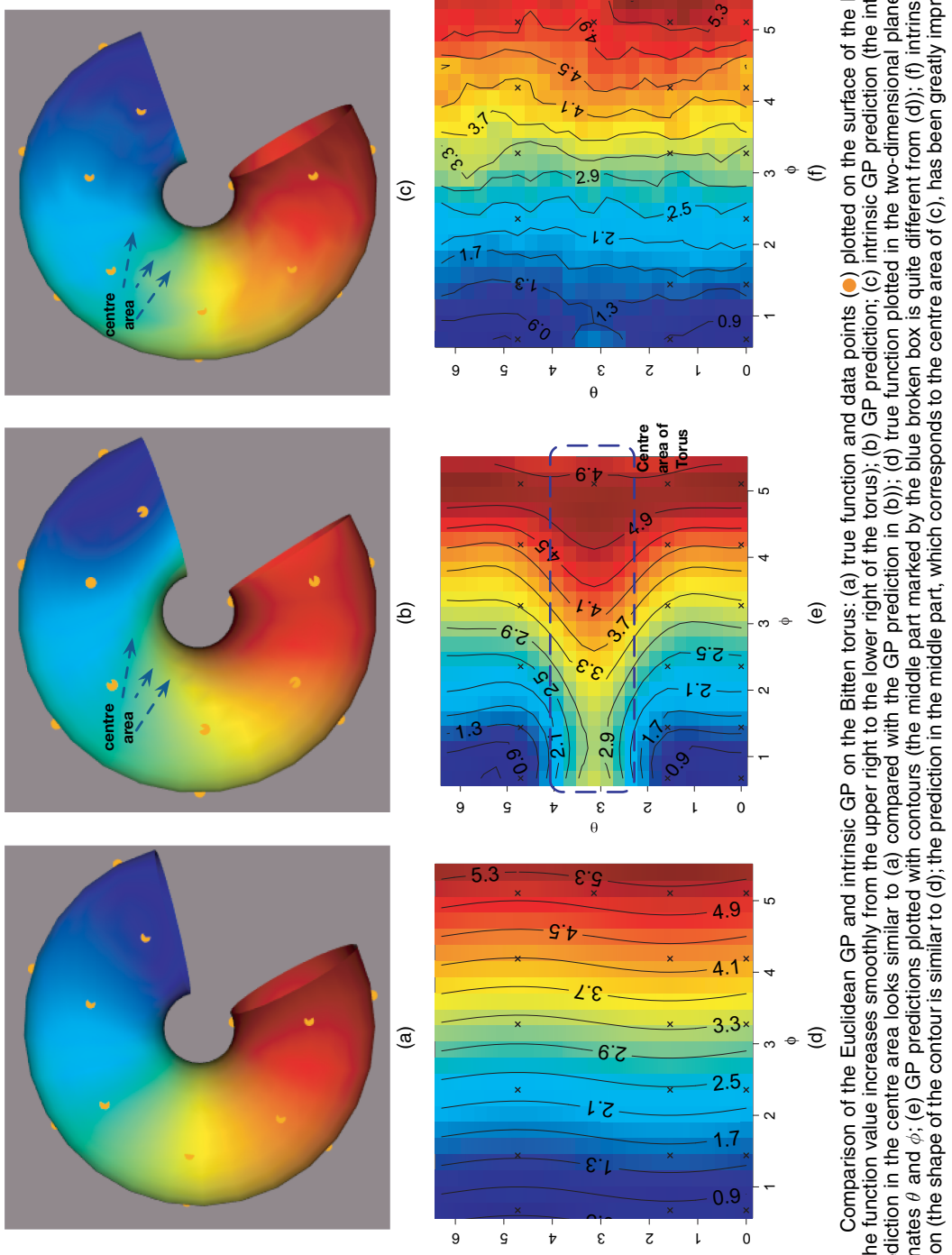

Fig. 8. Comparison of the Euclidean GP and intrinsic GP on the Bitten torus: (a) true function and data points (●) plotted on the surface of the Bitten torus (the function value increases smoothly from the upper right to the lower right of the torus); (b) GP prediction; (c) intrinsic GP prediction (the intrinsic GP prediction in the centre area looks similar to (a) compared with the GP prediction in (b)); (d) true function plotted in the two-dimensional plane with co-ordinates θ and ϕ; (e) GP predictions plotted with contours (the middle part marked by the blue broken box is quite different from (d)); (f) intrinsic GP prediction (the shape of the contour is similar to (d); the prediction in the middle part, which corresponds to the centre area of (c), has been greatly improved)

f is defined on the surface of the Bitten torus with

$$Y_i = f(x_i, y_i, z_i) + \epsilon_i,$$

where x_i, y_i and z_i are the co-ordinates of a point on the surface. More details on construction of the torus and the derivation of the metric tensor are shown in the on-line supplementary material. A demonstration of the BM on the Bitten torus is shown in Fig. 5(c).

We first applied a Euclidean GP to this example by using an RBF kernel in $\mathbb{R}^3$. The GP predictive mean is plotted in Fig. 8(b) with colour. Compared with the true function in Fig. 8(a), the GP predictive mean is brighter (the colour is similar to yellow) in the centre area. The Euclidean distance between a data point and a grid point in the centre area is smaller than the geodesic distance on the torus surface. The RBF kernel assigns bigger covariances between these points which makes the data point in the centre dominate the region.

By applying the intrinsic GP to these data, the BM sample paths can be simulated by using equation (9) with the metric tensor of the Bitten torus. In particular, BM on the Bitten torus can be modelled via the stochastic differential equations

$$\mathrm{d}\theta(t) = -\tfrac{1}{2} r^{-1} \sin(\theta)\{R + r\cos(\theta)\}^{-1}\mathrm{d}t + r^{-1}\mathrm{d}B_\theta(t), \tag{34}$$

$$\mathrm{d}\phi(t) = |\{R + r\cos(\theta)\}^{-1}|\mathrm{d}B_\phi(t). \tag{35}$$

A trace plot of a single BM sample path is shown in Fig. 5(c). Following the procedure in Section 2.1, the predictive mean of the intrinsic GP is plotted in Fig. 8(c). The intrinsic GP prediction in the centre area looks more similar to the true function in Fig. 8(a). Also the colour in the region near the lower bound is dark red, which is more similar to the true function compared with the GP prediction in Fig. 8(b).

For visualization convenience, we have also plotted the function in two dimensions of ϕ (the angle of the tube) and θ (the angle of the torus) in Fig. 8(d). The differences of the prediction in the centre area are clearer from the two-dimensional contour plot of the GP prediction in Fig. 8(e) and the intrinsic GP prediction in Fig. 8(f). In the two-dimensional contour plots, the distance between $\theta = 0$ and $\theta = 2\pi$ is 2π in R^2. However, given a fixed ϕ, $\theta = 0$ and $\theta = 2\pi$ represent the same point on the torus and the distance between them is 0. Therefore, methods such as Sampson and Guttorp's (1992) mapping the domain with a diffeomorphism to a regular region of R^k can lead to a big error for this case.

The performance of the Euclidean GP and intrinsic GP are compared by varying the noise with various signal-to-noise ratios. The values of the true function are perturbed by Gaussian noise (30 db and 10 db) with 50 replicates for each level of noise. For each of the replicates, different methods are applied to estimate the test function at some equally spaced grid points. The mean and standard deviation of the mean-squared error for these 50 replicates are reported in Table 5. The prediction of the intrinsic GP is significantly better at all levels of noise.

We have also carried out experiments by removing the data point from the centre of the torus. Details on more comparison results are also provided in the on-line supplementary material.

4. Application to chlorophyll data in the Aral sea

In this section, we consider an analysis of remotely sensed chlorophyll data at 485 locations in the Aral sea. The data are available from the `gamair` package (Wood, 2006) and are plotted in Fig. 9(a). The level of chlorophyll concentration is represented by the intensity of the colour. The

Table 5. Comparison of the root-mean-squared errors of predictive means of two methods on the Bitten torus†

Case	*Result for the following methods:*	
	Euclidean GP	*Intrinsic GP*
30 db	167.3 (11.7)	25.8 (11.6)
10 db	229.2 (93.4)	74.6 (32.7)

†The table shows the mean of root-mean-squared errors over 50 data sets. Values in parentheses show the standard deviation.

chlorophyll data from the satellite sensors are noisy and vary smoothly within the boundary but not across the gap corresponding to the isthmus of the peninsula. We applied different methods to estimate the spatial pattern of the chlorophyll density.

The *logarithm* of chlorophyll concentration is modelled as a function of the latitude and longitude co-ordinates of the measurement locations:

$$\text{chl}_i = f(\text{lon}_i, \text{lat}_i) + \epsilon_i,$$

where lon_i and lat_i are standardized by subtracting the mean.

To reduce the computation cost, the sparse intrinsic GP from Section 2.4 is applied. 42 inducing points are introduced that are equally spaced within the boundary of the Aral sea (represented by small triangles in Fig. 9(e)). The number of BM sample paths has been reduced from $485N_{\text{BM}}$ to $42N_{\text{BM}}$, where N_{BM} is 20000 in this example.

The predictive mean of the Euclidean GP is shown in Fig. 9(c). As expected, the Euclidean GP smooths across the isthmus of the central peninsula. Relatively high levels of chlorophyll concentration are estimated for the southern part of the eastern shore of the western basin of the sea, whereas all observations in this region have rather low concentrations. Similarly a decline in level of chlorophyll towards the southern half of the western shore of the eastern basin is estimated, which is different from the pattern of the data in the region. In contrast, the predictive mean using a sparse intrinsic GP does not produce these artefacts (see Fig. 9(e)) and tracks the data pattern better. The values of the predictive variance are plotted as a heat map in Fig. 1(a) in the on-line supplementary material. The level curves are orthogonal to the boundary in the north and east part of the Aral sea in Fig. 9(e). The inducing points are more sparse in the west part of the Aral sea. Approximation errors, due to insufficiently dense inducing points and/or Monte Carlo errors in approximating RBM with resampling, can lead to non-orthogonality in some cases.

These artefacts become even more pronounced when the coverage of the data is uneven. In Fig. 9(b) we removed most of the data points in the southern part of the western basin of the sea, and the same models are applied to this uneven data set. Fig. 9(d) shows the Euclidean GP extrapolation across the isthmus from the eastern basin of the sea. In contrast, the sparse intrinsic GP estimates as plotted in Fig. 9(f) do not seem to be affected by the data from the eastern side of the isthmus. The values of the predictive variance are plotted as a heat map in Fig. 1(b) in the supplementary material.

Since most of the data points in the southern part of the western basin of the sea have been removed, the values of the variance estimates have increased in this region. To compare with the

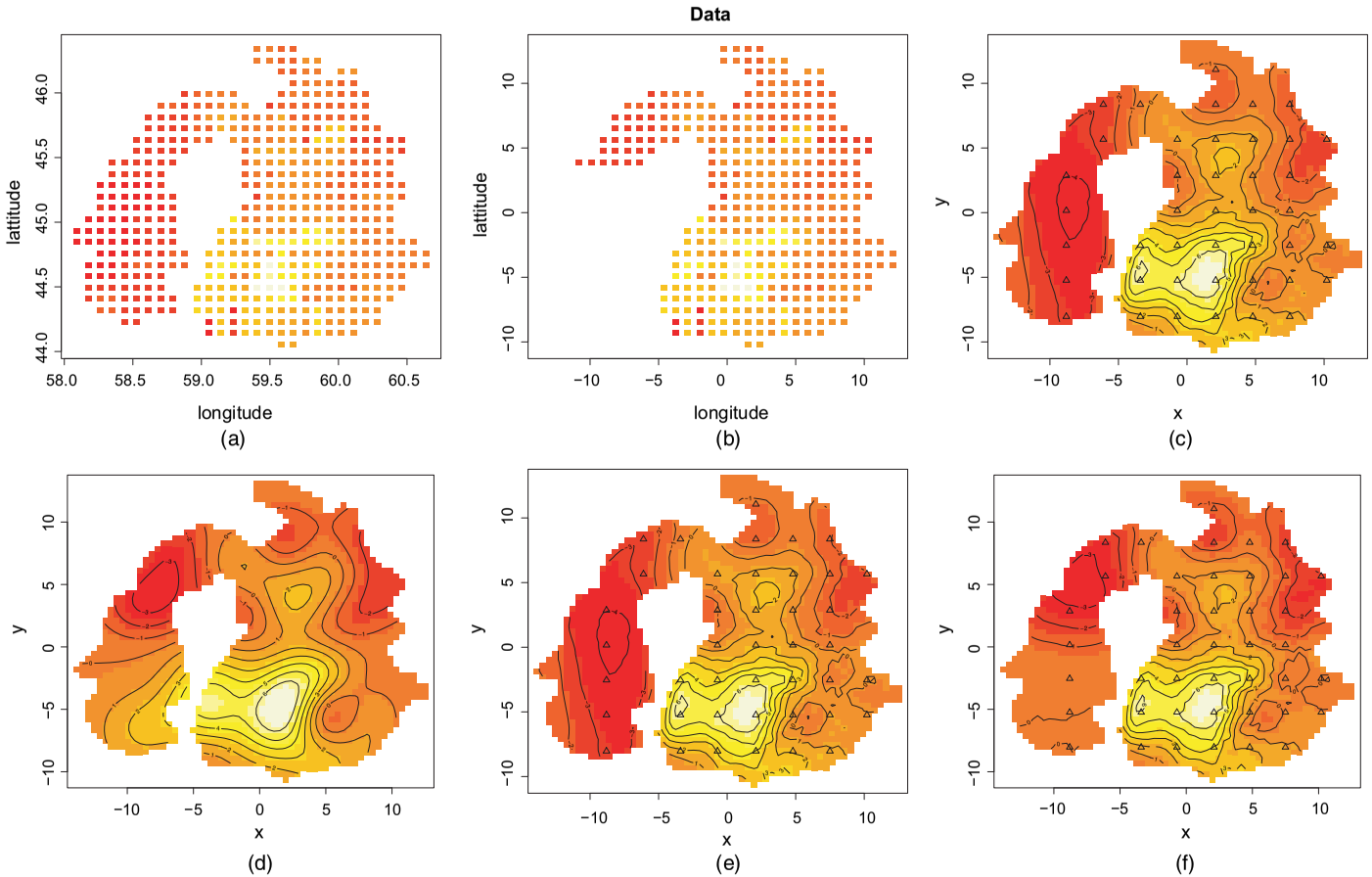

Fig. 9. Comparison of the intrinsic GP and a Euclidean GP for the chlorophyll data in the Aral sea: (a) remotely sensed chlorophyll data at 485 locations in the Aral sea (the data sets are available from the gamair package (Wood, 2006)); (b) chlorophyll data with the west basin removed; (c) GP prediction; (d) GP prediction with less data; (e) sparse intrinsic GP prediction; (f) sparse intrinsic GP prediction with less data (it is clear that the Euclidean GP smooths across the isthmus of the central peninsula in (c); the sparse intrinsic GP tracks the data pattern better; if we removed the data points in the southern part of the western basin of the sea in (b), these artefacts are exacerbated for the Euclidean GP prediction in (d); the sparse intrinsic GP prediction in (f) is not affected by the data from the eastern side of the isthmus

soap film approach, we divide the data sets into 10 equal size batches randomly and iteratively pick one batch as the training data to make prediction for the other nine batches. The mean root-mean-squared error RMSE for the intrinsic GP is 17.9 with standard deviation 0.71. The mean RMSE for the soap film method is 17.2 with standard deviation 1.72. The statistical test shows no significant difference between these two methods. However, the standard deviation of the RMSE for the intrinsic GP is much smaller.

5. Discussion

Our work proposes a novel class of intrinsic GPs on manifolds and complex constrained domains employing the equivalence relationship between heat kernels and the transition density of BM on manifolds. One of the key features of the intrinsic GP is to incorporate fully the intrinsic geometry of the spaces for inference while respecting the potentially complex boundary or interior constraints. To reduce the computational cost of simulating BM sample paths when the sample size is large, sparse intrinsic GPs are developed leveraging ideas from the literature on fast computation in GPs in Euclidean spaces. The results in Sections 3 and 4 indicate that an intrinsic GP achieves significant improvement over usual GPs. Although we did not conduct a formal asymptotic study for intrinsic GPs, with insights gained from the Euclidean GP with squared exponential kernel in the Euclidean space (see, for example, van der Vaart and van Zanten (2009)), we expect intrinsic GPs to yield posterior consistency with respect to appropriate neighbourhoods of the true regression function f_0. The focus of this paper has been on developing intrinsic GPs on manifolds with known metric tensors. There has been abundant interest in learning of unknown lower dimensional manifold structure in high dimensional data. Intrinsic GPs can be combined with these approaches for performing supervised learning on lower dimensional latent manifolds.

6. Code and supplementary material

R code implementation of the examples in Sections 3 and 4 are available on the GitHub repository `https://github.com/mu2013/Intrinsic-GP-on-complex-constrained-domain`. The on-line supplementary material includes a description on the choice of the sample and window sizes for estimating the heat kernel in $\mathbb{R}$ and $\mathbb{R}^2$. It also provides some details on the BM on the Swiss roll and the Bitten torus. A discussion on the differences between reflecting and resampling for the BM is also provided. The last section of the supplementary material is devoted to a discussion of the computational cost and algorithm complexity.

Acknowledgement

Lizhen Lin acknowledges support for this paper from National Science Foundation grants DMS CAREER 1654579 and IIS 1663870.

References

Aumentado-Armstrong, T. and Siddiqi, K. (2017) Stochastic heat kernel estimation on sampled manifolds. *Comput. Graph. Forum*, **36**, 131–138.

Berline, N., Getzler, E. and Vergne, M. (2003) *Heat Kernels and Dirac Operators*. Berlin: Springer.

Bhattacharya, A. and Dunson, D. (2010) Nonparametric Bayes regression and classification through mixtures of product kernels. *Baysn Anal.*, **9**, 145–164.

Burdzy, K., Chen, Z.-Q. and Sylvester, J. (2004) The heat equation and reflected Brownian motion in time-dependent domains. *Ann. Probab.*, **32**, 775–804.

Castíllo, I., Kerkyacharian, G. and Picard, D. (2014) Thomas Bayes' work on manifolds. *Probab. Theory Reltd Flds*, **158**, 665–710.
Guinness, J. and Fuentes, M. (2016) Isotropic covariance functions on spheres: some properties and modeling considerations. *J. Multiv. Anal.*, **143**, 143–152.
Hsu, E. P. (2008) A brief introduction to Brownian motion on a Riemannian manifold. *Lecture Notes*.
Hsu, P. (1984) Reflecting Brownian motion, boundary local time and the Neumann problem. *Dissertn Abstr. Int.* B, **45**.
Hsu, P. (1988) Brownian motion and Riemannian geometry. *Contemp. Math.*, **73**, 95–104.
Katzfuss, M. and Guinness, J. (2017) A general framework for Vecchia approximations of Gaussian processes. *Preprint arXiv:1708.06302*.
Kloeden, P. E. and Platen, E. (1992) Higher-order implicit strong numerical schemes for stochastic differential equations. *J. Statist. Phys.*, **66**, 283–314.
Lamberton, D. and Lapeyre, B. (2007) *Introduction to Stochastic Calculus Applied to Finance*. Boca Raton: CRC Press.
Lin, L., Mu, N., Chan, P. and Dunson, D. B. (2018) Extrinsic Gaussian processes for regression and classification on manifolds. *Baysn Anal.*, to be published.
Lions, P. L. and Sznitman, A. S. (1984) Stochastic differential equations with reflecting boundary conditions. *Communs Pure Appl. Math.*, **37**, 511–537.
Matheron, G. (1973) The intrinsic random functions and their applications. *Adv. Appl. Probab.*, **5**, 439–468.
Ozakin, A. and Gray, A. G. (2009) Submanifold density estimation. In *Proc. Neural Information Processing Systems* (eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta), pp. 1375–1382. Red Hook: Curran Associates.
Pelletier, B. (2005) Kernel density estimation on Riemannian manifolds. *Statist. Probab. Lett.*, **73**, 297–304.
Quiñonero-Candela, J. and Rasmussen, C. E. (2005) A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.*, **6**.
Quiñonero-Candela, J., Rasmussen, C. E. and Williams, C. K. (2007) Approximation methods for Gaussian process regression. In *Large-scale Kernel Machines* (eds L. Bottou, O. Chapelle, D. DeCoste and J. Weston), pp. 203–224. Cambridge: MIT Press.
Ramsay, T. (2002) Spline smoothing over difficult regions. *J. R. Statist. Soc.* B, **64**, 307–319.
Rasmussen, C. E. (2004) Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning* (eds O. Bousquet, U. Luxburg and G. Rätsch), pp. 63–71. Berlin: Springer.
Sampson, P. D. and Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Statist. Ass.*, **87**, 108–119.
Sangalli, L. M., Ramsay, J. O. and Ramsay, T. O. (2013) Spatial spline regression models. *J. R. Statist. Soc.* B, **75**, 681–703.
Schwaighofer, A. and Tresp, V. (2002) Transductive and inductive methods for approximate Gaussian process regression. In *Advances in Neural Information Processing Systems*, **15**, pp. 953–960.
Snelson, E. and Ghahramani, Z. (2006) Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18* (eds Y. Weiss, B. Schölkopf and J. C. Platt), pp. 1257–1264. Cambridge: MIT Press.
Titsias, M. K. (2009) Variational learning of inducing variables in sparse Gaussian processes. In *Proc. 12th Int. Conf. Artificial Intelligence and Statistics* (eds D. van Dyk and M. Welling), pp. 567–574.
Tosi, A., Hauberg, S., Vellido, A. and Lawrence, N. D. (2014) Metrics for probabilistic geometries. In *Proc. 30th Conf. Uncertainty in Artificial Intelligence*, pp. 800–808. Association for Uncertainty in Artificial Intelligence Press.
van der Vaart, A. W. and van Zanten, J. H. (2009) Adaptive bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, **37**, 2655–2675.
Wood, S. (2006) *Generalized Additive Models: an Introduction with R*, 1st edn. Boca Raton: Chapman and Hall–CRC.
Wood, S. N. (2001) mgcv: gams and generalized ridge regression for R. *R News*, **1**, 20–25.
Wood, S. N., Bravington, M. V. and Hedley, S. L. (2008) Soap film smoothing. *J. R. Statist. Soc.* B, **70**, 931–955.
Zhou, Y., Cai, W. and Hsu, E. P. (2017) Computation of local time of reflecting Brownian motion and probabilistic representation of the Neumann problem. *Communs Math. Sci.*, **15**, 237–259.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for "Intrinsic Gaussian processes on complex constrained domains"'.