NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

02-22

Spherical Poisson Point Process Intensity Function Modeling and Estimation with Measure Transport

Tin Lok James Ng and Andrew Zammit-Mangion

Spherical Poisson Point Process Intensity Function Modeling and Estimation with Measure Transport

Tin Lok James Ng[1], Andrew Zammit-Mangion[2]

[1]School of Computer Science and Statistics, Trinity College Dublin, Ireland
[2]School of Mathematics and Applied Statistics, University of Wollongong, Australia

Abstract

Recent years have seen an increased interest in the application of methods and techniques commonly associated with machine learning and artificial intelligence to spatial statistics. Here, in a celebration of the ten-year anniversary of the journal *Spatial Statistics*, we bring together normalizing flows, commonly used for density function estimation in machine learning, and spherical point processes, a topic of particular interest to the journal's readership, to present a new approach for modeling non-homogeneous Poisson process intensity functions on the sphere. The central idea of this framework is to build, and estimate, a flexible bijective map that transforms the underlying intensity function of interest on the sphere into a simpler, reference, intensity function, also on the sphere. Map estimation can be done efficiently using automatic differentiation and stochastic gradient descent, and uncertainty quantification can be done straightforwardly via nonparametric bootstrap. We investigate the viability of the proposed method in a simulation study, and illustrate its use in a proof-of-concept study where we model the intensity of cyclone events in the North Pacific Ocean. Our experiments reveal that normalizing flows present a flexible and straightforward way to model intensity functions on spheres, but that their potential to yield a good fit depends on the architecture of the bijective map, which can be difficult to establish in practice.

Keywords:
Exponential map, Maximum likelihood, Normalizing flows, Radial flows, Wrapping potential functions

1. Introduction

The journal *Spatial Statistics* has attracted contributions in a large range of topics related to the discipline since its inception ten years ago. One of the topics that has featured consistently during its lifetime is that on point processes: between 10% and 30% of the articles published in the journal every year contained the word 'point[-]process' or 'point pattern' in either the title, the abstract, or the keywords. On the other hand, other topics have only recently began to feature in submissions, notably those related to machine learning and artificial intelligence (ML & AI). For example, six articles contained the words 'neural network' or 'random forest' in the year 2021 (7% of all publications in 2021), more than in all the other years combined.

This trend is not unique to the journal *Spatial Statistics*, and reflects an increased adoption and awareness of modeling and inferential strategies that have been developed in the ML & AI community. While the predictive ability and the use of technological innovation of models and algorithms in ML & AI is undisputed, statisticians have a lot to contribute to this field, particularly in the area of uncertainty quantification. Largely for historical reasons, statisticians also tend to be involved in applications that are different from those in the ML & AI community, and are drivers of integrating these new techniques with more classical statistical approaches in areas such as official statistics, geophysics, and ecology (e.g., McDermott and Wikle, 2019; Schafer et al., 2020).

Articles that incorporate ML & AI in spatial statistics include those of Gerber and Nychka (2021) and Lenzi et al. (2021), who use neural networks for estimating parameters governing spatial processes; McDermott and Wikle (2019), who use echo state networks for the modeling and forecasting of spatio-temporal phenomena; Zammit-Mangion and Wikle (2020), who use convolution neural networks to model the dynamics of geophysical phenomena in a dynamic spatio-temporal model; and Sidén and Lindsten (2020); Li et al. (2020); Zammit-Mangion et al. (2021); Murakami et al. (2021), who use ideas and technologies in deep learning to model arbitrarily complex spatial covariances or data models in spatial applications. This list is by no means exhaustive, but further highlights the considerable recent awareness of the role ML & AI has in solving some of the pertinent challenges in spatial statistics in the last few years.

In this article we home in on a topic that today is commonly associated with ML & AI: normalizing flows, which find their origin in an optimization problem known as 'optimal transport'. In the simplest application of a normalizing flow, one expresses a *target density* (i.e., a density function we wish to model) $f_0(\cdot)$, via the change of

variables formula,

$$f_0(\boldsymbol{x}) = f_1(T(\boldsymbol{x}))|\det(\nabla T(\boldsymbol{x}))|, \quad \boldsymbol{x} \in \mathcal{X}, \tag{1}$$

where $f_1(\cdot)$ is a *reference density*, commonly a normal density function, ∇ is the gradient operator, $T(\cdot)$ is an arbitrarily complex, bijective, differentiable, map, and $\mathcal{X}$ is a subset of the Euclidean plane. In ML & AI, normalizing flows are conventionally used in several applications involving density-function estimation (e.g., Papamakarios et al., 2021) and variational Bayes (e.g., Rezende and Mohamed, 2015), however notions related to these flows, and optimal transport more generally, are also starting to make an appearance in spatial statistics. For example, Katzfuss and Schäfer (2021) use transport maps to model the joint distribution of a spatial process at a large number of locations; Ng and Zammit-Mangion (2020) use these maps on spatial coordinates to model non-homogeneous spatial Poisson point processes; and Chen et al. (2020) use a type of normalizing flow, coined the continuous-time normalizing flow, for modeling spatio-temporal point patterns from earthquake and epidemiological data. Often, uncertainty in the map is not quantified or assessed, but there are exceptions to this: Katzfuss and Schäfer (2021) model transport maps using Gaussian processes, while Ng and Zammit-Mangion (2020) present a nonparametric bootstrap approach to generate prediction intervals of the estimated intensity function at any spatial location.

In addition to point processes and ML & AI, at least one article every year since 2014 in *Spatial Statistics* was concerned with modeling on the sphere. Methods for modeling on the sphere are important for modeling processes at a global scale, however work in this area has attracted less attention than the Euclidean case. Research on the modeling of point processes on the sphere is even scarcer, and much of the literature and applications concerning point processes is on the plane (e.g., Dias et al., 2008; Adams et al., 2009; Miranda and Morettin, 2011; Illian et al., 2012; Zammit-Mangion et al., 2012), although there has been a marked increased interest in the spherical case in recent years (Robeson et al., 2014; Lawrence et al., 2016; Møller et al., 2018).

In recognition of the substantial and diverse contributions *Spatial Statistics* has made to the community in the past decade, in this work we propose a methodology that ties together several of the threads of research discussed above: an approach for modeling the intensity function of a non-homogeneous Poisson point process on the sphere using normalizing flows. The paper is structured as follows: Section 2 outlines background material related to measure transport on the plane and on the sphere; Section 3 exploits the notion of process densities in Poisson processes to develop

a measure transport approach to model intensity functions on spheres; Section 4 provides illustrations of the method on both simulated and application data; and Section 5 concludes.

2. Background

In this section we give some background on measure transport for density-function estimation, which is required for the intensity-function estimation procedure outlined in Section 3. We first focus on the case of density-function estimation on the Euclidean space in Section 2.1 before moving to the spherical case in Section 2.2.

2.1. Transport of Probability Measure on Euclidean Spaces

Given two probability measures $\mu_0(\cdot)$ and $\mu_1(\cdot)$ defined on spaces $\mathcal{X}$ and $\mathcal{Z}$, respectively, a transport map $T : \mathcal{X} \to \mathcal{Z}$ is said to push forward $\mu_0(\cdot)$ to $\mu_1(\cdot)$ if, for any Borel subset $B \subset \mathcal{Z}$,

$$\mu_1(B) = \mu_0(T^{-1}(B)), \tag{2}$$

where the inverse $T^{-1}(\cdot)$ is set valued; specifically, $T^{-1}(\boldsymbol{z}) = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{x}) = \boldsymbol{z}\}$. For an injective transport map $T(\cdot)$, (2) can be re-formulated as

$$\mu_1(T(A)) = \mu_0(A), \tag{3}$$

for any Borel subset $A \subset \mathcal{X}$. Suppose that $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$, and that the measures $\mu_0(\cdot), \mu_1(\cdot)$ are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, with densities $\mathrm{d}\mu_0(\boldsymbol{x})/\mathrm{d}\boldsymbol{x} = f_0(\boldsymbol{x})$ and $\mathrm{d}\mu_1(\boldsymbol{z})/\mathrm{d}\boldsymbol{z} = f_1(\boldsymbol{z})$, respectively. If the map $T(\cdot)$ is bijective with a differentiable inverse $T^{-1}(\cdot)$, we obtain the familiar change-of-variables formula shown in (1), which expresses a complicated probability density $f_0(\cdot)$ in terms of a simpler density $f_1(\cdot)$ and a transport map $T(\cdot)$.

A transport map allows for the simulation of random variables from complicated probability distributions. Suppose our interest is to draw a sample from the probability measure $\mu_0(\cdot)$. Given a simpler measure $\mu_1(\cdot)$ that is easy to simulate from and a transport map $T(\cdot)$ that satisfies (2), one can first simulate a sample $\boldsymbol{Z}$ from $\mu_1(\cdot)$, and then apply the inverse transport map $T^{-1}(\cdot)$ to $\boldsymbol{Z}$; the resulting sample $T^{-1}(\boldsymbol{Z})$ will then be distributed according to $\mu_0(\cdot)$. The inverse transport map $T^{-1}(\cdot)$ may not be analytically available in general and, in such cases, numerical procedures for evaluating the inverse map are required.

4

The measure transport approach presents an attractive way to estimate probability density functions as it only involves specifying a reference density $f_1(\cdot)$, which is typically chosen to be the density of the standard multivariate normal distribution, and a transport map $T(\cdot)$. Marzouk et al. (2016) discusses various strategies for parameterizing the transport map $T(\cdot)$, and more recent approaches typically parameterize $T(\cdot)$ using a neural network (e.g., Papamakarios et al., 2017; Huang et al., 2018; Kobyzev et al., 2020). The bijectivity and differentiability properties of the map $T(\cdot)$ are preserved under compositions. That is, given two bijective maps $T^{(1)}(\cdot), T^{(2)}(\cdot)$ with differentiable inverses, their composition $T^{(2)} \circ T^{(1)}(\cdot)$ remains bijective with a differentiable inverse, and hence there is no issue of "space folding." The Jacobian determinant of the resulting composition remains computationally tractable since, by the chain rule,

$$\det(\nabla(T^{(2)} \circ T^{(1)}(\boldsymbol{x}))) = \det(\nabla T^{(1)}(\boldsymbol{x}))\det(\nabla T^{(2)}(T^{(1)}(\boldsymbol{x}))), \quad \boldsymbol{x} \in \mathcal{X}.$$

These properties associated with taking compositions facilitate the construction of complex transformations via the composition of multiple simpler transformations. Thus, an approach often used is to define $T(\cdot)$ as a composite of multiple transformations, that is, as $T(\cdot) \equiv T^{(K)} \circ \cdots \circ T^{(1)}(\cdot)$, where $T^{(k)}(\cdot)$ transforms $\boldsymbol{z}^{(k-1)}$ into $\boldsymbol{z}^{(k)}$, with $\boldsymbol{z}^{(0)} \equiv \boldsymbol{x}$ and $\boldsymbol{z}^{(K)} \equiv \boldsymbol{z}$. This 'flow' of transformations leads to the class of maps known as normalizing flows which, as discussed in Section 1, are widely used in ML & AI. Here, the problem of density-function estimation is equivalent to that of estimating the unknown parameters of the transport maps $T^{(1)}(\cdot), \ldots, T^{(K)}(\cdot)$, which can be efficiently performed using automatic differentiation libraries, stochastic gradient descent and, with maps involving neural networks, graphics processing units.

2.2. Transport of Probability Measure on Spheres

While most of the literature on measure transport and normalizing flows to date has concentrated on Euclidean spaces, some recent work has focused on normalizing-flow techniques for density-function estimation on Riemannian manifolds such as spheres and tori (Gemici et al., 2016; Mathieu and Nickel, 2020; Rezende et al., 2020). Gemici et al. (2016) developed a general approach for probability density-function estimation for Riemannian manifolds by first projecting the manifold to the Euclidean space $\mathbb{R}^d$, applying conventional Euclidean normalizing flows in this space, and then projecting $\mathbb{R}^d$ back to a manifold that is not necessarily the same as the original manifold. However, if the manifolds are not diffeomorphic to $\mathbb{R}^d$, as in the case of spheres, singularities arise that can lead to numerical instabilities during parameter estimation. To circumvent this difficulty, normalizing flows tailored for

Riemannian manifolds have been developed; these include exponential map flows, recursive Möbius spline flows (Rezende et al., 2020), and Riemannian continuous normalizing flows (Mathieu and Nickel, 2020).

In this paper we focus on the exponential map flow proposed by Rezende et al. (2020), which is based on a mechanism for establishing density functions on the sphere first proposed by Sei (2013), which in turn builds on the theoretical results of McCann (2001). The core idea is to construct a transport map from the space of *wrapping potential functions*, $W(\mathbb{S}^{d-1})$, $d \geq 2$, which contains differentiable functions on the sphere that exhibit a property known as *c-convexity* (see Definition 3.1 in Sei, 2013, for a formal definition). Consider, for now, a single wrapping potential function $\phi(\cdot) \in W(\mathbb{S}^{d-1})$. A valid transport map $G_\phi : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ can be constructed by taking the *exponential map* of the gradient $\nabla\phi(\cdot) \in T_{\boldsymbol{x}}\mathbb{S}^{d-1}$ where, with a slight abuse of notation, $T_{\boldsymbol{x}}\mathbb{S}^{d-1}$ is the tangent space of $\mathbb{S}^{d-1}$ at $\boldsymbol{x}$. (An exponential map $\exp_{\boldsymbol{x}}(\vec{\boldsymbol{v}})$, $\vec{\boldsymbol{v}} \in T_{\boldsymbol{x}}\mathbb{S}^{d-1}$, not to be confused with the exponential operator, returns the end-position of a particle that starts off at $\boldsymbol{x}$ and travels for one time unit with velocity $\vec{\boldsymbol{v}}$ on $\mathbb{S}^{d-1}$.) Specifically, given a wrapping potential function $\phi(\cdot)$,

$$T(\boldsymbol{x}) = G_\phi(\boldsymbol{x}) = \exp_{\boldsymbol{x}}(\nabla\phi(\boldsymbol{x})), \quad \boldsymbol{x} \in \mathbb{S}^{d-1}, \tag{4}$$

is a valid transport map, in the sense that it is homeomorphic and locally diffeomorphic. Interestingly, while the exponential map $G_\phi(\cdot)$ is generally analytically intractable for arbitrary Riemannian manifolds, a closed form expression for $\exp_{\boldsymbol{x}} : T_{\boldsymbol{x}}\mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ in Equation (4) is available:

$$\exp_{\boldsymbol{x}}(\vec{\boldsymbol{v}}) = (\cos\|\vec{\boldsymbol{v}}\|)\boldsymbol{x} + (\sin\|\vec{\boldsymbol{v}}\|)(\vec{\boldsymbol{v}}/\|\vec{\boldsymbol{v}}\|), \quad \vec{\boldsymbol{v}} \in T_{\boldsymbol{x}}\mathbb{S}^{d-1}, \tag{5}$$

where $\|\cdot\|$ is the Euclidean norm. One can therefore readily construct a probability distribution on the sphere with density function

$$f_\phi(\boldsymbol{x}) \propto |\det(\nabla G_\phi(\boldsymbol{x}))|, \quad \boldsymbol{x} \in \mathbb{S}^{d-1},$$

where we have taken the uniform measure as the reference measure, and where the Jacobian is taken with respect to any orthonormal basis on $T_{\boldsymbol{x}}\mathbb{S}^{d-1}$ and $T_{G_\phi(\boldsymbol{x})}\mathbb{S}^{d-1}$. Simulating random variables from this probability distribution is straightforward and involves first sampling $\boldsymbol{Z} \sim \text{Unif}(\mathbb{S}^{d-1})$, the uniform distribution on $\mathbb{S}^{d-1}$, and then solving $\boldsymbol{X} = G_\phi^{-1}(\boldsymbol{Z})$. Efficient methods exist for drawing samples from $\text{Unif}(\mathbb{S}^{d-1})$; for example, one could first sample from the standard d-dimensional Gaussian distribution, and then normalize the sample so that each d-dimensional vector has a

magnitude of one. The equality $\boldsymbol{X} = G_\phi^{-1}(\boldsymbol{Z})$ can also be solved efficiently (McCann, 2001). The exponential map flow offers a simple, intuitive, and flexible approach to modeling density functions on $\mathbb{S}^{d-1}$, since one only needs to specify a wrapping potential function $\phi(\cdot)$ for its construction.

The space of wrapping potential functions $W(\mathbb{S}^{d-1})$ has the desirable property of closure under convex combinations. That is, if $\phi_i(\cdot) \in W(\mathbb{S}^{d-1})$, for $i = 1, \ldots, p$, then $\sum_{i=1}^p \eta_i \phi_i(\cdot) \in W(\mathbb{S}^{d-1})$ if $\eta_i > 0$ and $\sum_{i=1}^p \eta_i = 1$. This property facilitates the construction of flexible families of probability distributions via relatively simple functions $\{\phi_i(\cdot)\}$. One general approach to construct $\phi(\cdot) \in W(\mathbb{S}^{d-1})$ is described in Sei (2013): Let $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_p \in \mathbb{S}^{d-1}$, and $\phi_1(\cdot), \ldots, \phi_p(\cdot) \in W(\mathbb{S}^{d-1})$, and define

$$\phi(\boldsymbol{x}) \equiv \sum_{i=1}^p \eta_i \phi_i(d(\boldsymbol{x}, \boldsymbol{m}_i)), \quad \boldsymbol{x} \in \mathbb{S}^{d-1}, \tag{6}$$

where $d(\boldsymbol{x}, \boldsymbol{y})$ is the geodesic distance between $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}$, and $\eta_i > 0$, $\sum_{i=1}^p \eta_i = 1$. In some cases, a closed form expression for the resulting probability density function $f_\phi(\cdot)$ is then available, which facilitates practical implementation. Various tractable specifications of the function $\phi_i(\cdot)$ are proposed in Sei (2013) and Rezende et al. (2020). In this paper we adopt the radial flow proposed in Rezende et al. (2020) where $\phi_i(\boldsymbol{x}) = e^{\beta_i(\cos d(\boldsymbol{x}, \boldsymbol{m}_i) - 1)} / \beta_i$, and, hence,

$$\phi(\boldsymbol{x}) = \sum_{i=1}^p \frac{\eta_i}{\beta_i} e^{\beta_i(\cos d(\boldsymbol{x}, \boldsymbol{m}_i) - 1)}, \quad \boldsymbol{x} \in \mathbb{S}^{d-1}, \tag{7}$$

where, for $i = 1, \ldots, p$, $\beta_i > 0, \boldsymbol{m}_i \in \mathbb{S}^{d-1}$, and $\eta_i > 0$ $(\sum_{i=1}^p \eta_i = 1)$ are model parameters that need to be estimated.

To help visualize the action of measure transport on spheres we give a small illustration. Consider (7) with $p = \eta_1 = 1$, $\beta_1 = 20$, and $\boldsymbol{m}_1 = (1, 0, 0)'$. This function, depicted in the left panel of Figure 1, has a peak at $\boldsymbol{m}_1$ and decays radially from this point. The vector field $\nabla \phi(\cdot)$ and its norm are shown in the middle and right panels of Figure 1, respectively. To illustrate the behavior of the exponential map of $\nabla \phi(\cdot)$, we simulate a random sample $\boldsymbol{Z}$ comprising 5000 points on the angular sub-domain $[-\pi/6, \pi/6] \times [-\pi/6, \pi/6]$ radians, and compute the terminal destination of these points after a time unit on spherical trajectories dictated by $\nabla \phi(\cdot)$, by evaluating (5) for each of these points. The uniformly-distributed points, and the exponential map of these points, are shown in the left and middle panels of Figure 2, respectively. Note how the exponential map has the effect of pulling elements in $\boldsymbol{Z}$
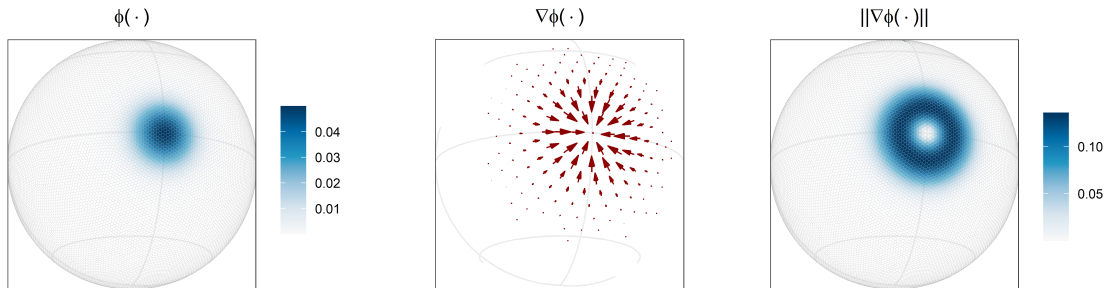
Figure 1: Left panel: The wrapping potential $\phi(\cdot)$ evaluated over an icosahedral Snyder equal area aperture 3 hexagon (ISEA3H) discrete global grid (DGG, Sahr et al., 2003) at resolution 7. Middle panel: The gradient field $\nabla\phi(\cdot)$ evaluated at the centroids of an ISEA3H DGG at resolution 4 with arrow sizes proportional to the vector norms. Right panel: The gradient norm $\|\nabla\phi(\cdot)\|$ evaluated over an ISEA3H DGG at resolution 7.
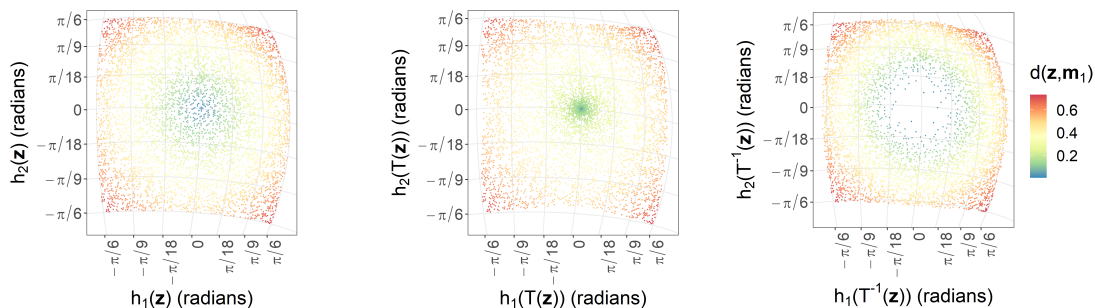

Figure 2: Left panel: A sample $\boldsymbol{Z}$ of 5000 points from a uniform distribution on the surface of the sphere in the region $[-\pi/6, \pi/6] \times [-\pi/6, \pi/6]$. Middle panel: The application of the exponential map $T(\cdot)$ to $\boldsymbol{Z}$. Right panel: The application of the inverse exponential map, $T^{-1}(\cdot)$, to $\boldsymbol{Z}$. In all panels, the colour is used to represent the geodesic distance of a point in the sample $\boldsymbol{Z}$ from $\boldsymbol{m}_1 = (1, 0, 0)'$, with blue, yellow, and red representing increasing distances, respectively, and the function $h(\cdot)$ maps from Cartesian to spherical coordinates.

toward $\boldsymbol{m}_1$. Now, recall that the exponential map is used to relate our target variable $\boldsymbol{X}$ to the reference variable $\boldsymbol{Z}$ via the mapping $T(\boldsymbol{X}) = \boldsymbol{Z}$, and that therefore a random sample of $\boldsymbol{X}$ is obtained by applying $T^{-1}(\cdot)$ to $\boldsymbol{Z}$. Intuitively, this inverse finds the initial departure location of points for which we know the terminal location. The effect is thus opposite to what we observe when applying $T(\cdot)$ to $\boldsymbol{Z}$, and the

inverse map causes elements in $\boldsymbol{Z}$ to diverge away from $\boldsymbol{m}_1$, as seen in the right panel of Figure 2.

We conclude this section by noting that, as in the Euclidean case, the flexibility of the resulting probability density function can be further improved by taking compositions of multiple transport maps:

$$G_\phi^{(K)} \circ \cdots G_\phi^{(1)}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{S}^{d-1}, \tag{8}$$

where each $G_\phi^{(k)}(\cdot), k = 1, \ldots, K$, has the form given in (4).

3. Intensity Function Modeling with Measure Transport

In this section we apply the theory outlined in Sections 2.1 and 2.2 to intensity function estimation of a non-homogeneous point process (NHPP).

A point process X on a domain $\mathcal{X}$ is a random finite subset of $\mathcal{X}$ with corresponding counting measure $N(\cdot)$, where for any subset $A \subset \mathcal{X}$, $N(A)$ denotes the number of points in X falling in A. For a homogeneous Poisson process with rate λ, $N(A)$ is a Poisson random variable with mean $\lambda|A|$, where $|A|$ is the area of the set A, and the random variables $N(A_1), N(A_2)$ are independent for disjoint regions A_1, A_2. An NHPP defined on $\mathcal{X}$ is a Poisson point process with varying intensity, which is fully characterized by its intensity function $\lambda : \mathcal{X} \to [0, \infty)$. NHPPs have mostly (but not exclusively) been modeled and fitted on Euclidean domains, that is, where $\mathcal{X} \subset \mathbb{R}^d$ and $d \geq 1$.

The inferential target when modeling NHPPs is the underlying intensity function. Nonparametric techniques, such as spline-based methods (Dias et al., 2008) and wavelet-based methods (Miranda and Morettin, 2011), which do not fix the functional form of the intensity function, are among the many popular approaches for intensity function estimation. However, these methods generally do not scale well with the number of observed points or the dimension d. Model-based approaches, such as those using a doubly-stochastic Poisson process (Møller et al., 1998) are also popular. However, inference with these models is typically computationally intensive, and often one needs to resort to approximate inferential methods for computations to remain tractable. Recently, Ng and Zammit-Mangion (2020) proposed a measure transport approach to model the intensity function of an NHPP on the Euclidean domain based on normalizing flows. The proposed approach was shown to be competitive with kernel-based methods and log-Gaussian Cox process models on simulation studies and a data application. In Section 3.1 we review this measure

9

transport approach for intensity function estimation on $\mathbb{R}^d$. Section 3.2 then extends the method to the spherical domain.

3.1. Intensity Function Modeling on $\mathbb{R}^d$ with Measure Transport

Consider a NHPP defined on $\mathcal{X} \subset \mathbb{R}^d$ with intensity function $\lambda(\cdot)$. The log-likelihood function of $\lambda(\cdot)$ with respect to a unit rate Poisson process can be written as (e.g., Møller et al., 1998),

$$\ell(\lambda; \{\boldsymbol{x}_i\}_{i=1}^n) = -\int_{\mathcal{X}} (\lambda(\boldsymbol{x}) - 1)\mathrm{d}\boldsymbol{x} + \sum_{i=1}^n \log \lambda(\boldsymbol{x}_i), \tag{9}$$

where $\{\boldsymbol{x}_i\}_{i=1}^n$ is a realization of the point process, and n is the number of points in the realization. Instead of optimizing the likelihood function in (9) directly, the core idea introduced by Ng and Zammit-Mangion (2020) is to model the so called "process density," $\tilde{\lambda}(\cdot) = \lambda(\cdot)/\mu_\lambda(\mathcal{X})$ using a measure transport approach, where $\mu_\lambda(\mathcal{X})$ is the integrated intensity defined as

$$\mu_\lambda(\mathcal{X}) = \int_{\mathcal{X}} \lambda(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

The quantity $\tilde{\lambda}(\cdot)$ can be regarded as a density function associated with the point process with respect to the Lebesgue measure on $\mathcal{X}$. The process density representation was introduced by Taddy and Kottas (2010); this representation is attractive as it allows the process density $\tilde{\lambda}(\cdot)$ to be modeled separately from $\mu_\lambda(\mathcal{X})$.

Ng and Zammit-Mangion (2020) model the process density $\tilde{\lambda}(\cdot)$ using the transport map $T: \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ need not be the same as $\mathcal{X}$. The map $T(\cdot)$ is then constructed from a class of normalizing flows known as neural autoregressive flows, which can approximate a large class of process densities arbitrarily well (Huang et al., 2018; Ng and Zammit-Mangion, 2020). Specifically, it can be shown that the class of neural autoregressive flows has the desirable "universal approximation property." Choosing a simple reference density function $\xi(\cdot)$ on $\mathcal{Z}$ such as the standard multivariate normal distribution if $\mathcal{Z} = \mathbb{R}^d$, and the uniform density if $\mathcal{Z}$ is bounded, the process density $\tilde{\lambda}(\cdot)$ can be expressed as

$$\tilde{\lambda}(\boldsymbol{x}; \Theta) = \xi(T(\boldsymbol{x}; \Theta))|\det(\nabla T(\boldsymbol{x}; \Theta))|, \quad \boldsymbol{x} \in \mathcal{X}.$$

where we have now explicitly denoted the dependence of $T(\cdot)$, and hence that of $\tilde{\lambda}(\cdot)$, on parameters Θ used in constructing the normalizing flow and that need to

be estimated. The log-likelihood function of the parameters is simply the log of the process density, and is given by

$$\ell(\Theta; \{\boldsymbol{x}_i\}_{i=1}^n) = \sum_{i=1}^n \left(\log \xi(T(\boldsymbol{x}_i; \Theta)) + \log |\det(\nabla T(\boldsymbol{x}_i; \Theta))| \right). \tag{10}$$

Maximum likelihood estimation of Θ, and hence of $\tilde{\lambda}(\cdot)$, reduces to numerical optimization of the log-likelihood function (10) with respect to Θ. Further, since $N(\mathcal{X})$ is a Poisson random variable with mean $\int_{\mathcal{X}} \lambda(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mu_\lambda(\mathcal{X})$, $\mu_\lambda(\mathcal{X})$ can be estimated by n, the number of points of the realized point process. Therefore, an estimate for the intensity function $\lambda(\cdot)$ is given by $\hat{\lambda}(\cdot) = n\hat{\tilde{\lambda}}(\cdot)$, where $\hat{\tilde{\lambda}}(\cdot) \equiv \tilde{\lambda}(\cdot; \hat{\Theta})$ is the estimated process density, and $\hat{\Theta}$ is the maximum likelihood estimate of Θ.

3.2. Intensity Function Modeling on $\mathbb{S}^{d-1}$ with Measure Transport

In this section we consider the problem of estimating the intensity function of a NHPP on $\mathbb{S}^{d-1} \equiv \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$ where $\|\cdot\|$ is the L^2 norm, for $d \geq 2$ but with a particular focus on $\mathbb{S}^2$. An NHPP on $\mathbb{S}^{d-1}$ is characterized by its intensity function $\lambda(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{S}^{d-1}$, where $N(A)$ is a Poisson random variable with mean $\int_A \lambda(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ for $A \subset \mathbb{S}^{d-1}$. We extend the measure transport approach to intensity function modeling on the Euclidean space described in Section 3.1 to the spherical domain by employing the normalizing flows described in Section 2.2.

Our objective is to estimate the intensity function $\lambda(\cdot)$ based on the realized points $\{\boldsymbol{x}_i\}_{i=1}^n$. Analogously to the Euclidean case, the log-likelihood function of $\lambda(\cdot)$ with respect to the unit rate Poisson process on the sphere can be written as

$$\ell(\lambda; \{\boldsymbol{x}_i\}_{i=1}^n) = -\int_{\mathbb{S}^{d-1}} (\lambda(\boldsymbol{x}) - 1)\mathrm{d}\boldsymbol{x} + \sum_{i=1}^n \log \lambda(\boldsymbol{x}_i). \tag{11}$$

As in Section 3.1, we define the process density as $\tilde{\lambda}(\cdot) = \lambda(\cdot)/\mu_\lambda(\mathbb{S}^{d-1})$, where $\mu_\lambda(\mathbb{S}^{d-1}) = \int_{\mathbb{S}^{d-1}} \lambda(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ is the integrated intensity, and separately estimate $\mu_\lambda(\mathbb{S}^{d-1})$ and $\tilde{\lambda}(\cdot)$. Since $N(\mathbb{S}^{d-1})$ is a Poisson random variable with mean $\int_{\mathbb{S}^{d-1}} \lambda(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mu_\lambda(\mathbb{S}^{d-1})$, $\mu_\lambda(\mathbb{S}^{d-1})$ can be estimated by n, the number of points in the realization. Since $\int_{\mathbb{S}^{d-1}} \tilde{\lambda}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$, $\tilde{\lambda}(\cdot)$ is a valid probability density on $\mathbb{S}^{d-1}$ with respect to the Lebesgue measure, and the measure transport framework described in Section 2.2 can be applied.

We apply the exponential map radial flow given by (4) and (7) to model the process

density $\tilde{\lambda}(\cdot)$. We use the uniform measure as the reference probability measure on $\mathbb{S}^{d-1}$; then, the log-likelihood function (10) simplifies to

$$\ell(\Theta; \{\boldsymbol{x}_i\}_{i=1}^n) = \sum_{i=1}^n \log |\det(\nabla G_\phi(\boldsymbol{x}_i; \Theta))| + \text{const.}, \tag{12}$$

where we have now explicitly denoted the dependence of $G_\phi(\cdot)$ on the model parameters $\Theta \equiv \{\beta_i^{(k)}, \boldsymbol{m}_i^{(k)}, \eta_i^{(k)} : k = 1, \ldots, K; i = 1, \ldots, p\}$ that construct a composition of K radial flows, where $G_\phi(\cdot) \equiv G_\phi^{(K)} \circ \cdots \circ G_\phi^{(1)}(\cdot)$ is the composition of K exponential maps with radial flow, and where the k-th map in the composition has model parameters $\{\beta_i^{(k)}, \boldsymbol{m}_i^{(k)}, \eta_i^{(k)}\}_{i=1}^p$. The simplification arises from our choice of the uniform measure as the reference measure, which leads to the first term in the summation of the log-likelihood function (10) to be constant (denoted as 'const.' in (12)). Estimating the process density $\tilde{\lambda}(\cdot)$ proceeds by (numerically) optimizing the log-likelihood function in (12) with respect to Θ.

To complete the specification of the model, we need to choose the number of maps in the composition, K, and the number of basis functions constructing each radial flow in (7), p. We found that a small value of p (e.g. $p = 1$ or $p = 2$) and a moderate to large value of K (between 20 and 40) led to good performance in our simulation experiments. This observation is consistent with much of the literature in deep learning (e.g., Eldan and Shamir, 2016; Raghu et al., 2017), where deep, narrow, neural networks are seen to have more representational efficiency than shallow, wide, networks. A theoretic rationale for choosing the optimal values of K and p is unfortunately not available. Their selection is analogous to the problem of choosing the optimal number of hidden layers and nodes per layer for deep neural network models, which is often solved by training a large number of models on training data, and testing them on (unseen) test data. We optimize (12) by using automatic differentiation libraries and stochastic gradient descent in *PyTorch* (Paszke et al., 2017). The estimated process density $\hat{\tilde{\lambda}}(\cdot)$ is then obtained from the estimated model parameters, and finally the estimated intensity function $\hat{\lambda}(\cdot)$ is simply given by $n\hat{\tilde{\lambda}}(\cdot)$.

To quantify the uncertainty on the estimated intensity function we proceed as in Ng and Zammit-Mangion (2020) and adopt a nonparametric bootstrapping approach (Efron, 1981). This approach requires the sampling of B bootstrap samples, which are generated as follows: First, the number of points $n_b, b = 1, \ldots, B$, is simulated from a Poisson distribution with rate parameter n. Second, for each bootstrap sample

$b = 1, \ldots, B$, n_b points from the observed points $\{\boldsymbol{x}_i\}_{i=1}^n$ are sampled with replacement. Third, B process densities, $\hat{\tilde{\lambda}}_b(\cdot), b = 1, \ldots, B$, are estimated using the above approach, yielding the estimated intensity functions $\hat{\lambda}_b(\cdot) = n_b \hat{\tilde{\lambda}}_b(\cdot), b = 1, \ldots, B$. Finally, we quantify the uncertainty at any point $\boldsymbol{x} \in \mathbb{S}^{d-1}$ through percentiles of $\{\hat{\lambda}_b(\boldsymbol{x}) : b = 1, \ldots, B\}$.

4. Simulation Studies and Application

In this section we assess the performance of our proposed method through simulation studies (Section 4.1) and an application where the aim is to model the intensity of cyclone events in the North Pacific Ocean (Section 4.2).

4.1. Simulation Experiment

In this section we carry out an experiment to assess the accuracy of the measure transport approach of Section 3.2 when estimating known intensity functions from point-process realizations, and also to assess the sensitivity of the estimates to the number of transport maps used in the composition of the normalizing flows.

In our simulation experiment, Poisson point-process realizations were simulated from the following intensity functions:

$$
\begin{align}
\lambda_1(\boldsymbol{x}) &= n f_{unif}(\boldsymbol{x}), \tag{13}\\
\lambda_2(\boldsymbol{x}) &= n f_{vmf}(\boldsymbol{x}; \tilde{\boldsymbol{m}}, \kappa), \tag{14}\\
\lambda_3(\boldsymbol{x}) &= \pi n f_{vmf}(\boldsymbol{x}; \tilde{\boldsymbol{m}}_1, \kappa_1) + (1 - \pi) n f_{vmf}(\boldsymbol{x}; \tilde{\boldsymbol{m}}_2, \kappa_2), \tag{15}\\
\lambda_4(\boldsymbol{x}) &= \pi_1 n f_{vmf}(\boldsymbol{x}; \tilde{\boldsymbol{m}}_1, \kappa_1) + \pi_2 n f_{vmf}(\boldsymbol{x}; \tilde{\boldsymbol{m}}_2, \kappa_2) + \\
&\quad (1 - \pi_1 - \pi_2) n f_{vmf}(\boldsymbol{x}; \tilde{\boldsymbol{m}}_3, \kappa_3), \tag{16}
\end{align}
$$

for $\boldsymbol{x} \in \mathbb{S}^2$, where $f_{unif}(\cdot)$ denotes the uniform density on $\mathbb{S}^2$; $f_{vmf}(\cdot; \tilde{\boldsymbol{m}}, \kappa)$ denotes the von Mises–Fisher distribution on $\mathbb{S}^2$ with location parameter $\tilde{\boldsymbol{m}}$ and concentration parameter κ; $\pi, \pi_1, \pi_2 \in [0, 1]$; and $\pi_1 + \pi_2 \in [0, 1]$. We set $n = 500$ for all intensity functions. For $\lambda_2(\cdot)$, we set $\tilde{\boldsymbol{m}} = (0, 0, 1)'$ and $\kappa = 100$. For $\lambda_3(\cdot)$, we set $\tilde{\boldsymbol{m}}_1 = (0, 1, 0)', \tilde{\boldsymbol{m}}_2 = (0, 0, 1)', \kappa_1 = \kappa_2 = 100$, and $\pi = 0.5$. For $\lambda_4(\cdot)$, we set $\tilde{\boldsymbol{m}}_1 = (1, 0, 0)', \tilde{\boldsymbol{m}}_2 = (0, 1, 0)', \tilde{\boldsymbol{m}}_3 = (0, 0, 1)', \kappa_1 = \kappa_2 = \kappa_3 = 100$, and $\pi_1 = \pi_2 = 1/3$. We considered a range of values for the number of transport maps K, and we set $p = 1$ for the radial flow in (7). We chose $f_{unif}(\cdot)$ as the reference density on $\mathbb{S}^2$. Rezende et al. (2020) showed, in the density estimation setting, that density functions constructed from a mixture of von Mises–Fisher density functions can be well estimated using exponential map radial flows. Therefore, the main
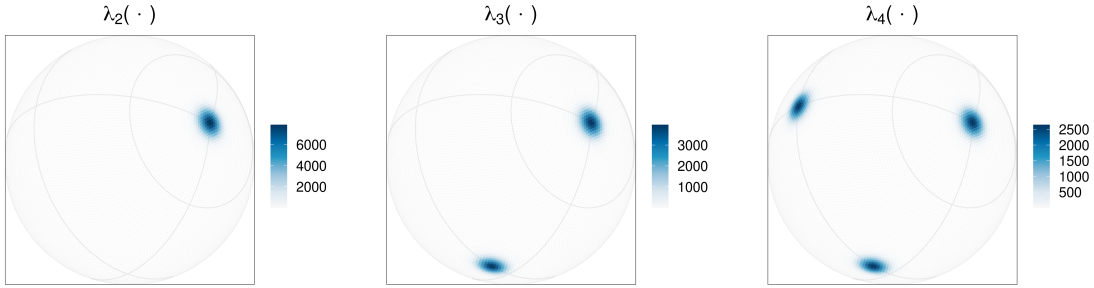
Figure 3: Intensity functions constructed via mixtures of von Mises–Fisher density functions that are used in the simulation experiment of Section 4.1. Left panel: The intensity function $\lambda_2(\cdot)$. Center panel: The intensity function $\lambda_3(\cdot)$. Right panel: The intensity function $\lambda_4(\cdot)$.

purpose of our simulation experiment is to assess the sensitivity of the method to different values of K, and to compare the quality of fits to that obtained using a log-Gaussian Cox process (LGCP) model-based approach to intensity function estimation. The intensity functions $\lambda_2(\cdot), \lambda_3(\cdot)$ and $\lambda_4(\cdot)$ are visualized in Figure 3.

For each intensity function $(\lambda_1(\cdot), \ldots, \lambda_4(\cdot))$, we generated 20 independent point-process realizations, and fit intensity functions to each of the realizations using transport maps with $K \in \{1, 5, 10, 20, 30, 40\}$. We assessed goodness of fit through the L^1 distances between the true and the estimated intensity functions, for each of the four cases. We also compared the goodness-of-fit metrics obtained from our proposed method to those from an LGCP model with a Matérn covariance function with smoothness parameter equal to one. The LGCP was fitted to the data using the package `inlabru` (Bachl et al., 2019) using an irregular finite-element tessellation of the sphere comprising approximately 2000 elements. This number of elements is sufficient given that $\lambda_1(\cdot), \ldots, \lambda_4(\cdot)$ vary relatively slowly (and not at all in the case of $\lambda_1(\cdot)$) spatially. The average and empirical standard deviations of the L^1 distances across the 20 realizations were computed for each case; the results are shown in Table 1.

From Table 1, we see that the proposed method considerably underperforms the LGCP when the true intensity function is uniform on $\mathbb{S}^2$ $(\lambda_1(\cdot))$, despite this also being the reference density for the normalizing flow. Further, the intensity-function estimate deteriorates as the number of compositions of transport maps, K, increases,

Table 1: Average and empirical standard deviation of the 20 L^1 distances between the true and the fitted intensity functions in the simulation experiment for each of $\lambda_1(\cdot), \ldots, \lambda_4(\cdot)$ as given in (13)–(16). Boldface denotes the best performance in an L^1 sense for each case.

Case	No. of compositions (K)	1	5	10	20	30	40	LGCP
$\lambda_1(\cdot)$	Average L^1 distance	10.63	23.45	24.43	25.37	28.11	28.67	**5.01**
	Stdev. of L^1 distance	2.07	2.87	1.88	1.89	4.40	3.17	1.89
$\lambda_2(\cdot)$	Average L^1 distance	68.83	52.56	21.52	**6.62**	16.11	22.41	11.61
	Stdev. of L^1 distance	2.08	1.82	2.39	1.48	7.12	6.03	1.78
$\lambda_3(\cdot)$	Average L^1 distance	69.70	56.45	34.81	**11.24**	14.51	15.67	13.90
	Stdev. of L^1 distance	1.00	3.55	4.92	2.56	2.49	2.21	1.78
$\lambda_4(\cdot)$	Average L^1 distance	65.93	52.01	40.96	22.82	**12.35**	16.45	14.71
	Stdev. of L^1 distance	1.05	2.79	2.52	3.60	2.06	2.35	1.49

and $K = 1$ (the least complex map) gives by far the best fit. This highlights a limitation of our approach with radial flow, which does not easily model the identity map due to its functional form; this lies in contrast with other warping approaches that allow the identity map to be easily inferred (e.g., Zammit-Mangion et al., 2021). On the other hand, when the true intensity functions are more variable ($\lambda_2(\cdot)$–$\lambda_4(\cdot)$), we find cases where transport maps begin to yield a better fit to the true underlying intensity function than a classic LGCP. In these examples, we find that $K = 20$–30 compositions (i.e., 20–30 radial flows) with $p = 1$ (i.e., one radial flow per compositional layer) give the best fits. These results show that, as is often the case with models constructed via composition such as deep learning models, the optimal number of layers to choose (K), and the width of each layer (p), is a non-trivial problem. This is especially the case here, where cross-validation (as is often done for model selection in this scenario) does not easily generalize to the point process case.

4.2. Proof of Concept: Application to Cyclone Data

In this section we apply our proposed method for spherical intensity function estimation to a dataset that contains location information on cyclones in the North Pacific Ocean. The dataset was published by the United States National Hurricane Center.[1] The dataset contains six-hourly information on the location, maximum winds, and central pressure of all known tropical and subtropical cyclones (1049 in

---

[1] https://www.nhc.noaa.gov/data/hurdat/hurdat2-nepac-1949-2020-043021a.txt

total). Cyclones are cause for significant infrastructure damage and financial loss (Pielke et al., 2008), and the modeling of cyclone prevalence is important for risk management and insurance purposes (e.g., Crabbe et al., 2008).

Here, we focus on the end locations of the cyclones. Based on the results in Section 4.1, we set the number of compositions to $K = 30$ and let $p = 1$. Optimization of (12) proved to be difficult in this case, with final estimates sensitive to the initial parameter settings. We hence used a "committee of networks" approach; see for example, Bishop (1995, Section 9.6) and Perrone and Cooper (1995). This strategy, where one trains several models with random initializations (in our case 50) and then averages their outputs, is often adopted when training highly parameterized deep networks that have multi-modal objective functions. Following model fitting, we applied the nonparametric bootstrap resampling methodology described in Section 3.2 with $B = 50$ to obtain prediction intervals for the intensity function (without using a committee of networks in each re-fitting stage). The estimated intensity function, along with the end locations of the cyclones, is shown in the center panel of Figure 4. The empirical 10 and 90 percentiles of the bootstrap distribution of the intensity function are shown in the left and right panels of Figure 4, respectively. One can add interpretation to the inferred normalizing flow by visualizing the action of the transport maps through the transformed locations of the cyclones at different stages of the map composition. We show six of these stages in Figure 5. Note how the transformed locations become increasingly uniformly distributed on the sphere as they are propagated through the network of compositions, as expected.

We conclude this proof of concept with a note on computation. Fitting these highly parameterized models formed through composition is surprisingly quick and computationally efficient. Recall that the objective function (12) is simply a sum of n independent quantities that can all be computed efficiently in parallel using high-dimensional arrays. Notably, no inverses are needed when evaluating the objective function or its gradient, and the model scales well with both data size n and the dimension d. In this example, evaluating the objective function and the gradient with respect to all the parameters required approximately 0.1 s, and we therefore only needed 10 s for fitting the model with 100 gradient steps. Further, fitting for the committee of networks and bootstrapping could be trivially parallelized, so that intensity-function estimation and uncertainty quantification could in principle be done in well under a minute on a multicore platform. The high computational efficiency also makes it possible to evaluate fits from different architectures with ease. These results are encouraging, and highlight measure transport as a viable
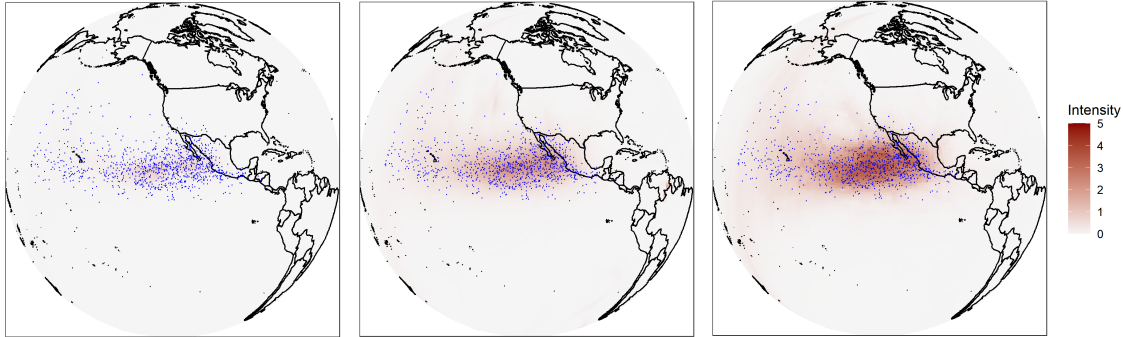
Figure 4: Intensity function (red surface; events per 10,000 square km) and the observed locations of cyclones (blue dots). Left panel: The empirical 10 percentile of the bootstrap distribution of the estimated intensity function based on $B = 50$ replicates. Center panel: The estimated intensity function. Right panel: Same as left panel, but for the 90 percentile.

alternative to conventional LGCP modeling when, for example, the spatial point process intensity function is considerably inhomogeneous or when one is operating in a higher-dimensional space (for example, in the case of 3D-space-time point processes).

We provide code for reproducing the results in this section at `https://github.com/andrewzm/SpherePP`.

5. Conclusion

This paper develops a general approach to the problem of intensity-function estimation of a non-homogeneous Poisson process on the sphere. The proposed approach leverages the framework of normalizing flows by using compositions of transport maps to model the unknown intensity function. The method is highly flexible and makes minimal assumptions on the target intensity function, which can be made arbitrary complex by increasing the number of maps in the composition and the number of basis functions in each radial flow. We illustrate the viability of our proposed approach in simulations studies, and its use in a proof of concept study where we model cyclone end locations in the North Pacific Ocean. Uncertainty quantification on the intensity function is done using nonparametric bootstrap.

It is not surprising that normalizing flows work well for point process intensity function estimation given their success in density estimation in several ML & AI applications (e.g., Rezende et al., 2020). Our approach also inherits their limita-
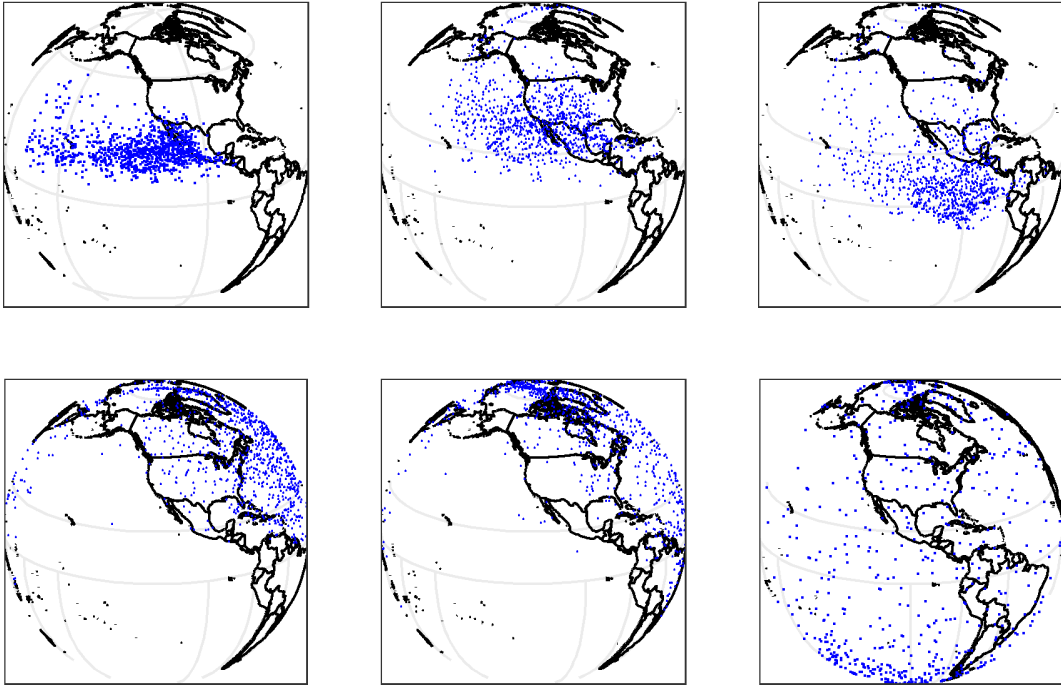
17

Figure 5: Top-left panel: The end locations of the cyclones used in this proof of concept. Top-center panel to bottom-right panel (left to right, top to bottom): The transformed locations after applying the map $G_\phi^{(k)} \circ \cdots \circ G_\phi^{(1)}(\cdot)$ to the locations, for $k = 6, 12, 18, 24, 30$, respectively. Recall that $K = 30$ for this model and that the reference density is uniform on $\mathbb{S}^2$.

tions. Specifically, choosing an optimal number of maps and number of radial flows is a challenging problem. Further, optimizing the objective function (12) requires the specification of suitable initial parameter values, mini-batch size and gradient step size when doing stochastic gradient descent; these settings, which can be difficult to tune in practice, where circumvented in our case through a committee of networks approach. These difficulties are a direct consequence of choosing a highly flexible model for the intensity function, and our preliminary results indicate that this approach only begins to pay dividends when the underlying intensity function is considerably non-homogeneous.

Several extensions of this work are possible. First, although we have considered

the nonparametric bootstrap for uncertainty quantification, one may also place prior distributions on the unknown parameters, and then implement stochastic variational Bayes as was done, for example, by Zammit-Mangion et al. (2021). Prior distributions may be helpful in regularizing the optimization problem. Second, while we have applied the exponential map radial flows in this paper, other normalizing flows on the spheres can also be considered, which might be more suitable for modeling point processes typically encountered in spatial statistics. Third, the incorporation of covariate information will be important for many practical applications, and an extension of our measure transport approach that allows for covariates is an interesting avenue of future research. Fourth, while we consider the setting of inhomogeneous Poisson process on $\mathbb{S}^2$ this work, the general normalizing flows framework is applicable to higher dimensional spheres. In higher dimensional settings, the normalizing flow approach will have significant computational advantages over LGCPs. Finally, theoretical properties of normalizing flows on Euclidean domains have been investigated where it has been shown that under mild conditions arbitrary probability distributions and intensity functions can be approximated arbitrarily well using normalizing flows with relatively simple transformations (Huang et al., 2018; Jaini et al., 2019; Ng and Zammit-Mangion, 2020). It is desirable to investigate whether normalizing flows on spherical and other non-Euclidean domains share such nice properties.

Acknowledgements

Declarations of Interest

Declarations of interest: none.

References

Adams, R. P., Murray, I., and MacKay, D. J. C. (2009), "Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY: Association for Computing Machinery, pp. 9–16.

Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019), "inlabru: an R package for Bayesian spatial modelling from ecological survey data," *Methods in Ecology and Evolution*, 10, 760–766.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford, UK: Oxford University Press.

Chen, R. T., Amos, B., and Nickel, M. (2020), "Neural spatio-temporal point processes," *arXiv preprint arXiv:2011.04583*.

Crabbe, M. J. C., Walker, E. L. L., and Stephenson, D. B. (2008), "The impact of weather and climate extremes on coral growth," in *Climate Extremes and Society*, eds. Diaz, H. F. and Murnane, R. J., Cambridge, UK: Cambridge University Press, p. 165–188.

Dias, R., Ferreira, C., and Garcia, N. (2008), "Penalized maximum likelihood estimation for a function of the intensity of a Poisson point process," *Statistical Inference for Stochastic Processes*, 11, 11–34.

Efron, B. (1981), "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods," *Biometrika*, 68, 589–599.

Eldan, R. and Shamir, O. (2016), "The power of depth for feedforward neural networks," in *Proceedings of the 29th Annual Conference on Learning Theory*, eds. Feldman, V., Rakhlin, A., and Shamir, O., New York, NY: PMLR, vol. 49 of *Proceedings of Machine Learning Research*, pp. 907–940.

Gemici, M. C., Rezende, D., and Mohamed, S. (2016), "Normalizing flows on Riemannian manifolds," *arXiv preprint arXiv:1611.02304*.

Gerber, F. and Nychka, D. (2021), "Fast covariance parameter estimation of spatial Gaussian process models using neural networks," *Stat*, 10, e382.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. (2018), "Neural autoregressive flows," in *Proceedings of the 35th International Conference on Machine Learning*, eds. Dy, J. and Krause, A., Stockholmsmässan, Stockholm, Sweden: PMLR, vol. 80 of *Proceedings of Machine Learning Research*, pp. 2078–2087.

Illian, J., Sorbye, S., and Rue, H. (2012), "A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA)," *Annals of Applied Statistics*, 6, 1499–1530.

Jaini, P., Selby, K. A., and Yu, Y. (2019), "Sum-of-squares polynomial flow," in *Proceedings of the 36th International Conference on Machine Learning*, eds. Chaudhuri, K. and Salakhutdinov, R., PMLR, vol. 97 of *Proceedings of Machine Learning Research*, pp. 3009–3018.

Katzfuss, M. and Schäfer, F. (2021), "Scalable Bayesian transport maps for high-dimensional non-Gaussian spatial fields," *arXiv preprint arXiv:2108.04211*.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020), "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 3964–3979.

Lawrence, T., Baddeley, A., Milne, R. K., and Nair, G. (2016), "Point pattern analysis on a region of a sphere," *Stat*, 5, 144–157.

Lenzi, A., Bessac, J., Rudi, J., and Stein, M. L. (2021), "Neural networks for parameter estimation in intractable models," *arXiv preprint arXiv:2107.14346*.

Li, Y., Sun, Y., and Reich, B. J. (2020), "DeepKriging: spatially dependent deep neural networks for spatial prediction," *arXiv preprint arXiv:2007.11972*.

Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A. (2016), "Sampling via Measure Transport: An Introduction," in *Handbook of Uncertainty Quantification*, eds. Ghanem, R., Higdon, D., and Owhadi, H., Cham, Switzerland: Springer International Publishing, pp. 785–825.

Mathieu, E. and Nickel, M. (2020), "Riemannian continuous normalizing flows," in *Advances in Neural Information Processing Systems 33*, eds. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., Red Hook, NY, USA: Curran Associates, Inc., pp. 2503–2515.

McCann, R. (2001), "Polar factorization of maps on Riemannian manifolds," *Geometric & Functional Analysis GAFA*, 11, 589–608.

McDermott, P. L. and Wikle, C. K. (2019), "Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data," *Entropy*, 21, 184.

Miranda, J. and Morettin, P. (2011), "Estimation of the Intensity of Non-homogeneous Point Processes via Wavelets," *Annals of the Institute of Statistical Mathematics*, 63, 1221–1246.

Møller, J., Nielsen, M., Porcu, E., and Rubak, E. (2018), "Determinantal point process models on the sphere," *Bernoulli*, 24, 1171–1201.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998), "Log Gaussian Cox processes," *Scandinavian Journal of Statistics*, 25, 451–482.

Murakami, D., Kajita, M., Kajita, S., and Matsui, T. (2021), "Compositionally-warped additive mixed modeling for a wide variety of non-Gaussian spatial data," *Spatial Statistics*, 43, 100520.

Ng, T. L. J. and Zammit-Mangion, A. (2020), "Non-Homogeneous Poisson process intensity modeling and estimation using measure transport," *arXiv preprint arXiv:2007.00248*.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021), "Normalizing flows for probabilistic modeling and inference," *Journal of Machine Learning Research*, 22, 1–64.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017), "Masked autoregressive flow for density estimation," in *Advances in Neural Information Processing Systems 30*, eds. Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., Red Hook, NY, USA: Curran Associates Inc., p. 2335–2344.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017), "Automatic Differentiation in PyTorch," Online: `https://openreview.net/group?id=NIPS.cc/2017/Workshop/Autodiff`.

Perrone, M. P. and Cooper, L. N. (1995), "When networks disagree: Ensemble methods for hybrid neural networks," in *How We Learn; How We Remember: Toward an Understanding of Brain and Neural Systems*, ed. Cooper, L. N., Singapore: World Scientific, pp. 342–358.

Pielke, R., Gratz, J., Landsea, C., Collins, D., Saunders, M., and Musulin, R. (2008), "Normalized hurricane damage in the United States: 1900–2005," *Natural Hazards Review*, 9, 29–42.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. (2017), "On the expressive power of deep neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, eds. Precup, D. and Teh, Y. W., PMLR, Sydney, Australia, vol. 70 of *Proceedings of Machine Learning Research*, pp. 2847–2854.

Rezende, D. and Mohamed, S. (2015), "Variational inference with normalizing Flows," in *Proceedings of the 32nd International Conference on Machine Learning*, eds. Bach, F. and Blei, D., Lille, France: PMLR, vol. 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538.

Rezende, D. J., Papamakarios, G., Racaniere, S., Albergo, M., Kanwar, G., Shanahan, P., and Cranmer, K. (2020), "Normalizing flows on tori and spheres," in *Proceedings of the 37th International Conference on Machine Learning*, eds. Daumé III, H. and Singh, A., PMLR, vol. 119 of *Proceedings of Machine Learning Research*, pp. 8083–8092.

Robeson, S. M., Li, A., and Huang, C. (2014), "Point-pattern analysis on the sphere," *Spatial Statistics*, 10, 76–86.

Sahr, K., White, D., and Kimerling, A. J. (2003), "Geodesic discrete global grid systems," *Cartography and Geographic Information Science*, 30, 121–134.

Schafer, T. L., Wikle, C. K., and Hooten, M. B. (2020), "Bayesian inverse reinforcement learning for collective animal movement," *arXiv preprint arXiv:2009.04003*.

Sei, T. (2013), "A Jacobian inequality for gradient maps on the sphere and its application to directional statistics," *Communications in Statistics – Theory and Methods*, 42, 2525–2542.

Sidén, P. and Lindsten, F. (2020), "Deep Gaussian Markov random fields," in *Proceedings of the 37th International Conference on Machine Learning*, eds. Daumé III, H. and Singh, A., PMLR, vol. 119 of *Proceedings of Machine Learning Research*, pp. 8916–8926.

Taddy, M. and Kottas, A. (2010), "Mixture modeling for marked Poisson processes," *Bayesian Analysis*, 7, 335–362.

Zammit-Mangion, A., Dewar, M., Kadirkamanathan, V., and Sanguinetti, G. (2012), "Point process modelling of the Afghan War Diary," *Proceedings of the National Academy of Sciences*, 109, 12414–12419.

Zammit-Mangion, A., Ng, T. L. G., Vu, Q., and Filippone, M. (2021), "Deep compositional spatial models," *Journal of the American Statistical Association*, doi: 10.1080/01621459.2021.1887741.

Zammit-Mangion, A. and Wikle, C. K. (2020), "Deep integro-difference equation models for spatio-temporal forecasting," *Spatial Statistics*, 37, 100408.