Nonparametric Fisher Geometry with Application to Density Estimation

Babak Shahbaba
Statistics
UC Irvine

Shiwei Lan
Mathematical and Statistical Sciences
Arizona State University

Jeffrey D. Streets
Mathematics
UC Irvine

Andrew J. Holbrook*
Biostatistics
UCLA

Abstract

It is well known that the Fisher information induces a Riemannian geometry on parametric families of probability density functions. Following recent work, we consider the nonparametric generalization of the Fisher geometry. The resulting nonparametric Fisher geometry is shown to be equivalent to a familiar, albeit infinite-dimensional, geometric object—the sphere. By shifting focus away from density functions and toward *square-root* density functions, one may calculate theoretical quantities of interest with ease. More importantly, the sphere of square-root densities is much more computationally tractable. As discussed here, this insight leads to a novel Bayesian nonparametric density estimation model.

1 INTRODUCTION

The Fisher information—and the geometry it induces—has been one of the unequivocal success stories of geometry in statistics. Building on recent work, we extend the Fisher geometry beyond parametric statistical models and show that the resulting geometry is equivalent to that of the infinite-dimensional sphere. The purpose of this paper is to bring attention to this new perspective and to demonstrate its theoretical and methodological consequences. As an application, we introduce the χ^2-process density prior, a flexible nonparametric model for Bayesian density estimation that admits fast computation while requiring minimal assumptions.

The Fisher information matrix is canonical in statistics: it is rooted in information theory (Gourieroux and Monfort, 1995); it appears in Jeffrey's prior of Bayesian analysis (Jeffreys, 1946); and it plays a central role in Bayesian and Frequentist asymptotics (Le Cam, 2012). Fisher advocated the importance of the information matrix in maximum likelihood estimation (Fisher, 1925). Fisher's student, Rao, was the first to place the information matrix in a differential geometric context (Rao, 1945). Since then, the differential geometric implications for parametric statistical models have been the subject of extensive inquiry (Amari and Nagaoka, 2007). Recently, a number of researchers have drawn connections between the Fisher geometry and the geometry of the infinite sphere (Srivastava, Jermyn, and Joshi, 2007; Chen, Streets, and Shahbaba, 2015; Itoh and Satoh, 2015; Kurtek and Bharath, 2015; Srivastava and Klassen, 2016; Peter, Rangarajan, and Moyou, 2017). Much of this work has been in the area of shape analysis and has focused on using the Fisher geometry to measure distance between probability densities. Bayesian uses for the nonparametric Fisher geometry were featured in (Chen, Streets, and Shahbaba, 2015), where Bayesian variational inference was accomplished by minimizing the Fisher distance, and in (Kurtek and Bharath, 2015), where the nonparametric Fisher geometry was used for sensitivity analysis of Bayesian models. Here, we focus on fully Bayesian nonparametric inference, including the generation of posterior samples using Hamiltonian Monte Carlo (HMC). In contrast to recent research, the geodesics associated with the nonparametric Fisher geometry are used to efficiently explore the MCMC state space and *not* to measure or minimize the distance between density functions.

This paper, and other recent research in the Fisher geometry, builds on the sub-field of square-root density estimation. (Pinheiro and Vidakovic, 1997) used a wavelet basis to estimate the square-root density by effectively fitting the curve and then normalizing a sparse collection of wavelet coefficients, and (Müller and Vidakovic, 1998) introduced a Bayesian follow-up to this work. Recently, (Hong and Gao, 2016) used Riemannian geometry to fit a square-root density model, but did not

*Corresponding Author: aholbroo@g.ucla.edu

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

make any connections to the Fisher geometry. More recently (Peter, Rangarajan, and Moyou, 2017) performed square-root density estimation for object recognition using minimum description length as fitting criterion and used the nonparametric Fisher geometry to obtain a closed-form expression of this criterion.

In this paper, we focus on the application of the nonparametric Fisher geometry to Bayesian inference for probability densities. While the density function is the object of interest, we instead model the square-root density function, that is, the function the square of which integrates to unity. We take a Bayesian nonparametric approach and endow the square-root density with a Gaussian process (GP) prior (Williams and Rasmussen, 1996) multiplied by a Dirac measure limiting its support to the infinite-dimensional sphere. In order to maintain this restriction, it is useful to use the Karhunen-Loève (K-L) expansion (Wang, 2008) of the GP prior as opposed to its kernel representation. Every GP with bounded second moment may be represented in terms of the eigenfunction expansion of its covariance operator, but this (the K-L) expansion is only explicitly known for a few classes of GPs (Wang, 2008). Still, the K-L expansion has seen much recent success in the realm of Bayesian inverse problems (Dashti and Stuart, 2013; Cotter et al., 2013) and has been featured in infinite-dimensional HMC and infinite manifold HMC (∞-mHMC) (Beskos et al., 2016). The proposed application of the K-L expansion to model the square-root density is unprecedented and offers a probabilistic interpretation to the use of basis expansions for density estimation.

Due to the orthonormality of the eigenfunction basis, the restriction to the (uncountably) infinite-dimensional sphere translates to a restriction to the (countably) infinite-dimensional sphere for the eigenvalues of the GP. Then, following the precedent set in (Beskos et al., 2016), the K-L expansion is truncated and the object of inference is reduced to the posterior distribution of a finite number of K-L coefficients restricted to a finite sphere. This computation is quick and easy using spherical HMC (Lan, Zhou, and Shahbaba, 2014). Thanks to the basis representation, computational complexity scales linearly with the number of data points, as opposed the cubic rate of the GP density sampler (Murray, MacKay, and Adams, 2009). Moreover, we show that—in the square-root density estimation context—spherical HMC corresponds to Riemannian HMC in the infinite-dimensional limit.

Squaring the GP square-root density prior gives a χ^2-process (cf. Rabier and Genz, 2014) density prior. We illustrate the use of this prior for a number of problems. The model is flexible and its posterior draws provide plausible realizations of the uncertainty inherent in the density estimation problem. Besides a recent application to Bayesian quadrature (Gunter et al., 2014), we are unaware of statistical applications for the χ^2-process and are therefore pleased to present its novel application to Bayesian density estimation. In that sense, our method can be considered as an alternative to Dirichlet Process Mixture Models (DPMM), which are commonly used for nonparametric Bayesian estimation. DPMMs convolve the Dirichlet process with a smooth distribution, in effect constructing an infinite mixture model (Antoniak, 1974). More recently, (Murray, MacKay, and Adams, 2009) proposed a new method, called Gaussian Process Density Sampler (GPDS), offering a similar amount of flexibility as the DPMM but having an arguably simpler framework. Nonetheless, inference for DPMM requires an advanced Gibbs sampling routine (Neal, 2000), and inference for the GPDS requires exchange sampling to handle the unit-integral restriction on the GP model (Murray, MacKay, and Adams, 2009). In contrast, the model we propose here can be computed using generic spherical HMC (Lan, Zhou, and Shahbaba, 2014) or geodesic Monte Carlo (Byrne and Girolami, 2013) algorithms.

In summary, the contributions of this paper are as follows:

- we review a nonparametric generalization of the Fisher geometry and show its relationship to the infinite-dimensional (L^2) sphere, the space of square-root density functions;
- we derive the geodesics on the L^2 sphere and use these geodesics to formalize the relationship between Riemannian HMC and infinite-dimensional spherical HMC;
- focusing on Bayesian nonparametric density estimation, we demonstrate the practical benefits to modeling the *square-root* density function. The resulting χ^2-process density prior performs well for a variety of problems and is efficiently computed using spherical HMC.

The rest of the paper is organized in the following way. In Section 2 we review the parametric Fisher geometry, present a nonparametric extension of the Fisher geometry, and derive key results by relating this geometry to the infinite-dimensional sphere. Section 3 presents the χ^2-process density prior along with some necessary tools, such as the Karhunen-Loève expansion. In Section 4, we discuss efficient Bayesian inference for the model and relate Riemannian HMC to infinite-dimensional spherical HMC. Section 5 relates our method to the Cox process Cox, 1955. Empirical results are presented in Section 6.

Finally, in Section 7 we discuss model limitations and possible extensions. All proofs are placed in the Supplement.

2 THE NONPARAMETRIC FISHER GEOMETRY

2.1 THE PARAMETRIC FISHER GEOMETRY

Given data x in domain $\mathcal{D}$, it is often useful to specify a probabilistic model $S = \{p_\theta = p(x,\theta) \,|\, \theta = \big[\theta^1, \ldots, \theta^p\big]\}$, where θ is a vector parameterizing the model and taking values in the continuous parameter space Θ. Then at any point $\theta \in \Theta$, the Fisher information is the expectation of the negative log-likelihood Hessian:

$$\mathcal{I}(\theta) = -\mathrm{E}_x\Big(\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^T}\Big) = -\int_{\mathcal{D}} \frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^T}\, p(x|\theta)\, \mu(dx)\,,$$

where $\ell(\theta) = \log p(x|\theta)$. The Fisher information encodes second-order functional information about $\ell(\theta)$. This fact explains the use of the Fisher information as a gradient preconditioning matrix in both (the Frequentist) Fisher scoring (Longford, 1987) and (the Bayesian) Riemannian HMC (Girolami and Calderhead, 2011). The Fisher information may also be written as the expected outer product of the score vector $\partial \log p(x|\theta)/\partial\theta$:

$$\begin{aligned}\mathcal{I}(\theta) &= \mathrm{E}_x\Big(\big(\frac{\partial \ell(\theta)}{\partial\theta}\big)\big(\frac{\partial \ell(\theta)}{\partial\theta}\big)^T\Big) \\ &= \int_{\mathcal{D}} \big(\frac{\partial \ell(\theta)}{\partial\theta}\big)\big(\frac{\partial \ell(\theta)}{\partial\theta}\big)^T\, p(x|\theta)\, \mu(dx)\,.\end{aligned}$$

The Fisher information is symmetric positive definite at any point $\theta \in \Theta$. Taking note of this fact, Rao (1945) interpreted the Fisher information matrix as a Riemannian metric tensor, i.e. a smoothly varying, symmetric positive definite matrix defined over the parameter space Θ. In this way, the Fisher information matrix induces a Riemannian metric $g_\theta(\cdot,\cdot)$ over Θ satisfying

$$g_\theta(\ell_i, \ell_j) = \mathcal{I}_{ij}(\theta)\,, \quad \text{and} \quad g_\theta(\psi,\phi) = \sum_{i,j} \psi^i \phi^j \mathcal{I}_{ij}(\theta)$$

for $\ell_i = \partial\ell(\theta)/\partial\theta_i$, $\psi = \sum_{k=1}^p \psi^k \ell_k$ and $\phi = \sum_{k=1}^p \phi^k \ell_k$. Hence, the Fisher information may be thought of as inducing a non-trival geometry on the otherwise Euclidean parameter space Θ. There has been much inquiry into the nature of the parametric Fisher geometry. Efron used the Fisher geometry to prove the second-order efficiency of the MLE for exponential family models (Efron, 1978), and Amari and Nagaoka (2007) has constructed a body of work around the Fisher geometry and its dual connections. More recently, Girolami and Calderhead (2011) successfully used the Fisher geometry to guide the Hamiltonian flow of their Riemannian HMC. In this paper, we take another tack by generalizing the notion of the Fisher geometry to nonparametric models.

2.2 Beyond parametric models

We consider probability distributions over smooth manifolds $\mathcal{D}$, of which $\mathcal{D} \cong \mathbb{R}^d$ is a special case. Having fixed a background measure μ, let

$$\mathcal{P} := \Big\{ p : \mathcal{D} \to \mathbb{R} \mid p \geq 0, \int_{\mathcal{D}} p(x)\, \mu(dx) = 1 \Big\}$$

be the space of probability density functions over $\mathcal{D}$. That is, $\mathcal{P}$ is the set of Radon-Nikodym derivatives of probability measures that are absolutely continuous with respect to μ. The following construction is agnostic to whether μ is the Lebesgue measure over $\mathcal{D} = \mathbb{R}^d$ or the Hausdorff measure over a general Riemannian manifold $\mathcal{D} = \mathcal{M}$.

We deal with the space $\mathcal{P}$ and do *not* fix a parametric model. Instead we give $\mathcal{P}$ the structure of an infinite dimensional (formal) Riemannian manifold. First, we think of it as a smooth manifold. Observe that for a given $p \in \mathcal{P}$, the tangent space can be identified with

$$T_p\mathcal{P} := \Big\{ \phi \in C^\infty(\mathcal{D}) \mid \int_{\mathcal{D}} \phi(x)\, \mu(dx) = 0 \Big\}\,.$$

This identification arises when one differentiates the unit measure condition on probability density functions. That is, for a smooth curve $p_t : (-\epsilon, \epsilon) \to \mathcal{P}$ satisfying $dp_t/dt|_{t=0} = \phi$, we have

$$\begin{aligned}0 &= \frac{d}{dt}\int_{\mathcal{D}} p_t(x)\, \mu(dx)\big|_{t=0} = \int_{\mathcal{D}} \frac{dp_t}{dt}(x)\, \mu(dx) \\ &= \int_{\mathcal{D}} \phi(x)\, \mu(dx)\,.\end{aligned}$$

Now that we have a smooth manifold and an associated tangent space, we may define a Riemannian metric, i.e. a smoothly varying, symmetric, non-degenerate, bilinear function $g(\cdot,\cdot)_p : T_p\mathcal{P} \times T_p\mathcal{P} \to \{0\} \cup \mathbb{R}^+$. Riemannian metrics are useful for developing a notion of distance on a manifold that does not depend on any embedding in Euclidean space. One may define uncountably many metrics on a general manifold, but we are interested in a generalization of the parametric Fisher information metric.

Definition 1. *Given $\mathcal{D}$, the* nonparametric Fisher information metric on $\mathcal{P}(\mathcal{D})$ *is (Srivastava, Jermyn, and*

Joshi, 2007; Srivastava and Klassen, 2016)

$$g_F(\phi,\psi)_p := \int_{\mathcal{D}} \frac{\phi(x)\psi(x)}{p(x)}\mu(dx). \tag{1}$$

This metric is a consistent generalization of the parametric Fisher information metric. To see this, consider the parametric model $p(x|\theta)$, with θ as a vector. Then each element θ_i of θ defines a curve $\Theta_i \to \mathcal{P}$, where Θ_i is a slice of Θ, and

$$\begin{aligned}\mathcal{I}_{ij}(\theta) &= \int_{\mathcal{D}} \ell_i \ell_j \, p(x|\theta)\mu(dx) \\ &= \int_{\mathcal{D}} \frac{p_i(x|\theta)}{p(x|\theta)} \frac{p_j(x|\theta)}{p(x|\theta)} \, p(x|\theta)\mu(dx) \\ &= \int_{\mathcal{D}} \frac{p_i(x|\theta)p_j(x|\theta)}{p(x|\theta)}\mu(dx)\,.\end{aligned}$$

Here, we have adopted the shorthand $p_i(x|\theta) = \partial p(x|\theta)/\partial\theta_i$. Expressed in a more invariant fashion, interpreting a model as a map $\theta : \Theta \to \mathcal{P}$, one has that the parametric Fisher metric is induced by the nonparameteric Fisher metric, i.e.

$$\theta^* g_F = g_\theta.$$

In what follows we make a nontrivial change of variables suggested by this geometric picture which provides various theoretical and computational simplifications. In particular, for various reasons the manifold $\mathcal{P}$ equipped with Riemannian metric (1) is not particularly easy to deal with. In order to calculate geometric quantities of interest (e.g. geodesics, distances), we shift focus to the L^2 unit sphere, i.e. the space of square-root density functions

$$\mathcal{Q} := \left\{ q : \mathcal{D} \to \mathbb{R} \mid \int_{\mathcal{D}} q(x)^2 \, \mu(dx) = 1 \right\}.$$

This space, which is identified with $\mathcal{P}$ by a simple transformation indicated below, provides a much simpler backdrop for calculations. This infinite-dimensional L^2 sphere is a surprisingly familiar object. Its tangent spaces and geodesics are formally the exact same as those of the finite dimensional sphere $\mathcal{S}^{n-1}$, the only difference being the replacement of the Euclidean inner product with the integral inner product of L^2:

$$\langle f, h\rangle_{L^2} = \int_{\mathcal{D}} f(x)h(x)\,\mu(dx)\,.$$

Remarkably, this simpler space is isometric to the space of density functions equipped with the nonparameteric Fisher metric defined above. See the supplementary file for more information along with some basic results.

As we will see below, not only is the L^2 sphere $\mathcal{Q}$ more theoretically tractable, it also turns out to be more computationally tractable. In the following sections, we take advantage of these two kinds of tractability to construct a Bayesian nonparametric model on $\mathcal{Q}$ and use it for an application in density estimation.

3 THE CHI-SQUARE PROCESS PRIOR

In this section, we transition from the theoretical to the methodological aspects of the nonparametric Fisher geometry. We find that the square-root representation $q = \sqrt{p}$ is of use practically as well as theoretically. Here we focus on its natural application for density estimation and show that Bayesian density estimation can be much easier when one shifts focus to the sphere of square-root densities.

Suppose we want to attribute a smooth density function to observed data $x_1, \ldots, x_n$ on finite domain $\mathcal{D} \subset \mathbb{R}^d$ and recall the definitions (from Section 2) of the space of density functions and the space of square-root density functions:

$$\begin{aligned}\mathcal{P} :=& \quad \left\{p : \mathcal{D} \to \mathbb{R} \mid p \geq 0, \textstyle\int_{\mathcal{D}} p(x)\,\mu(dx) = 1\right\} \\ \mathcal{Q} :=& \quad \left\{q : \mathcal{D} \to \mathbb{R} \mid \textstyle\int_{\mathcal{D}} q(x)^2\,\mu(dx) = 1\right\},\end{aligned}$$

respectively. We want to find a suitable element $p(\cdot) \in \mathcal{P}(\mathcal{D})$, the space of functions over domain $\mathcal{D}$. Although this space contains the functions of interest, we opt to deal with the space $\mathcal{Q}$ of square-root densities instead. As stated in the prior section, $\mathcal{Q}$ is the unit sphere in the infinite-dimensional Hilbert space $L^2(\mathcal{D})$. We model the square-root density with a GP prior (or a Gaussian measure in L^2) multiplied by the Dirac measure restricting the function to the unit sphere:

$$q \sim \mathcal{GP} \times \delta_q(\mathcal{Q})\,. \tag{2}$$

It turns out that it is much easier to enforce the constraint given by Dirac measure $\delta_q(\mathcal{Q})$ than it is to enforce the corresponding constraint $\delta_p(\mathcal{P})$ (as is done for the GPDS). To do so, however, we do not represent the GP prior using its kernel representation as is commonly done in the literature. We opt instead to represent q in terms of the eigenvalues and orthonormal eigenfunctions of its covariance operator.

3.1 KARHUNEN-LOÈVE REPRESENTATION

In order to tractably enforce the constraint $\delta_q(\mathcal{Q})$ in (2), it is helpful to write q as a function (or linear sum of functions) for which we know the values of both

$$\int_{\mathcal{D}} q(x)\mu(dx) \quad \text{and} \quad \int_{\mathcal{D}} q(x)^2\mu(dx)\,.$$

This condition is satisfied by representing random function q as a linear combination of orthonormal basis functions. The K-L representation (Wang, 2008) provides a canonical way of doing so and thus links our fully probabilistic approach to other square-root density methods that rely on a basis (Pinheiro and Vidakovic, 1997; Müller and Vidakovic, 1998; Hong and Gao, 2016). Let $u(\cdot) \sim \mathcal{GP}(0, K(\cdot))$ be a mean zero Gaussian process over domain $\mathcal{D}$ with covariance operator $K(\cdot)$. Then u admits a K-L expansion of the form

$$u(\cdot) = \sum_{i=1}^{\infty} u_i \, \phi_i(\cdot), \qquad u_i \overset{ind}{\sim} N(0, \lambda_i^2), \tag{3}$$

where the λ_is and the ϕ_is are respectively the eigenvalues and eigenfunctions of operator K. That is to say, they satisfy

$$K(\phi_i)(x') = \int k(x, x')\phi_i(x)\mu(dx) = \lambda_i \phi_i(x')$$

where $k(\cdot, \cdot)$ is the usual covariance kernel. The eigenvalues are decreasing and their sum-of-squares is finite: $\lambda_{i+1} < \lambda_i$, $\sum_{i=1}^{\infty} \lambda_i^2 < \infty$. Finally, the eigenfunctions form an orthonormal basis of L^2:

$$\int \phi_i(x)\phi_j(x)\mu(dx) = 0, \quad \text{and} \quad \int \phi_i^2(x)\mu(dx) = 1\,.$$

In this paper, we model q as belonging to the Matérn class of GPs. For the Matérn class, a closed-form orthonormal basis may be obtained from the eigenfunctions of the Laplacian (Chung, 2013; Beskos et al., 2016). The covariance operator is given by

$$K = \sigma^2(\alpha - \Delta)^{-s}\,, \tag{4}$$

where α and σ^2 are positively constrained scale parameters, s is a smoothness parameter, and Δ is the Laplacian $\sum_{i=1}^{d} \partial_i^2$. The eigenvalues and eigenfunctions corresponding to this covariance operator depend on the area and dimensionality of domain $\mathcal{D}$ and are presented in Section 6 below. It should be noted that the decision to use the Matérn class is entirely dictated by ease of computation and does not preclude other classes of GP from being used in future applications.

3.2 THE MODEL

The proposed density model is Bayesian nonparametric, i.e. we place a prior distribution on a set of functions and eschew a restrictive parametric form. Given data $x = (x_1, \cdots, x_N) \in \mathcal{D}$, we obtain a posterior distribution, which is itself a distribution over the same set of functions and is absolutely continuous with respect to the specified prior distribution. As stated above, the prior $\pi(q)$ on square-root density $q \in \mathcal{Q}$ is a GP multiplied by the Dirac measure on the L^2 sphere. Following (3), the prior for q and the likelihood of the data x given q are given by

$$\pi(q) \propto \delta_q(\mathcal{Q}) \prod_{i=1}^{\infty} \exp\big(-q_i^2/(2\lambda_i^2)\big)\,,$$
$$\pi(x|q) = \prod_{n=1}^{N} q^2(x_n)\,,$$

since q is the square-root density. This prior can also be interpreted as arising from an infinite-dimensional Bingham distribution on the coefficients (Dryden, 2005). The posterior distribution on q is then given by

$$\pi(q|x) = \frac{\pi(x|q)\,\pi(q)}{\int_{\mathcal{Q}} \pi(x|q)\,\pi(q)\,dq} \propto \pi(q) \prod_{n=1}^{N} q^2(x_n)\,.$$

Suppressing the Dirac measure, the log-posterior given data $x_{1:N}$ may be written in terms of the K-L expansion of q:

$$\begin{aligned}\log \pi(q|x) &\propto \sum_{n=1}^{N} \log q(x_n)^2 - \frac{1}{2}\sum_{i=1}^{\infty} q_i^2/\lambda_i^2 \\ &= 2\sum_{n=1}^{N} \log|q(x_n)| - \frac{1}{2}\sum_{i=1}^{\infty} q_i^2/\lambda_i^2 \\ &= 2\sum_{n=1}^{N} \log\Big|\sum_{i=1}^{\infty} q_i\phi_i(x_n)\Big| - \frac{1}{2}\sum_{i=1}^{\infty} q_i^2/\lambda_i^2\,.\end{aligned}$$

By modelling the square-root density q with a GP prior, we model the density function p with a χ^2-process prior. Modeling the density p as a χ^2-process, we automatically enforce the non-negativity requirement for probability density functions. On the other hand, χ^2-processes are not restricted to have unit integrals. We therefore rely on a geometric HMC inference scheme to restrict proposals to the L^2 sphere. This is discussed in details in Section 4.

4 INFERENCE

Inference for the χ^2-process density model is relatively straightforward and amenable to advanced HMC methods. In Section 4.1, we show that, in this context, infinite-dimensional spherical HMC is equivalent to Riemannian HMC using the parametric Fisher information. In practice, we follow Beskos et al. (2016) and truncate the K-L expansion of the GP square-root density prior for an integer I using truncation operator T_I:

$$T_I\big(q(x)\big) = T_I\Big(\sum_{i=0}^{\infty} q_i\,\phi_i(x)\Big) = \sum_{i=0}^{I} q_i\,\phi_i(x)\,.$$

Due to the orthonormality of the basis ϕ_i, the unit integral constraint on $T_I(q)^2$ translates directly to a spherical constraint on the random coefficients $q^I = (q_0, \cdots, q_I)$. That is,

$$1 = \int_{\mathcal{D}} T_I\big(q(x)\big)^2 \mu(dx) = \int_{\mathcal{D}} \Big(\sum_{i=0}^{I} q_i \, \phi_i(x) \Big)^2 \mu(dx)$$
$$= \sum_{i=0}^{I} q_i^2 \int \phi_i(x)^2 \mu(dx) = \sum_{i=0}^{I} q_i^2$$

where the penultimate equality is given by the orthogonality of the basis elements and the last equality is on account of the basis elements being normal. Thus, inference can be performed over the coefficients q^I by using spherical HMC (Lan, Zhou, and Shahbaba, 2014) on the sphere $\mathcal{S}^I$. Both of these methods augment the state space with an auxiliary velocity variable v (satisfying $v^T q^I = 0$) and simulate from a Hamiltonian system by splitting (Shahbaba et al., 2014) the Hamiltonian of interest (H) into two Hamiltonians ($H^1 + H^2$):

$$H(q^I, v) = -\log \pi(q^I) + \frac{1}{2} G(q^I) + \frac{1}{2} v^T v$$
$$H^1(q^I, v) = -\log \pi(q^I) + \frac{1}{2} G(q^I)$$
$$H^2(q^I, v) = \frac{1}{2} v^T v \, .$$

Here π is the posterior distribution and G is the canonical Riemann tensor for the sphere (Lan, Zhou, and Shahbaba, 2014). Simulating from H^1 involves a small perturbation of the velocity by the gradient of H^1 with respect to q^I; simulating H^2 involves moving along the sphere's geodesics in the direction v. This last fact is relevant to the discussion of the following section.

The computational bottlenecks for both HMC and spherical HMC are the likelihood evaluations (within the accept-reject step) and the gradient evaluations (within the discretized trajectory). For our model, both likelihood and gradient evaluations require a single summation over N terms, each a simple function of the N observations individually. Thus, the complexity is linear in the number of data points ($O(N)$). This is orders faster than the $O(N^3)$ computations required to perform inference for the GPDS (Murray, MacKay, and Adams, 2009). However, in a big data setting, even linear complexity might prove too costly. In such case, we recommend performing these summations using a binary reduction on a GPU with $O(log_2(N))$ complexity (Holbrook et al., 2020).

4.1 INFERENCE IN THE LIMIT

We note that both spherical HMC uses geodesic flows on the finite dimensional sphere to propose new Markov chain states. Since these flows are formally equivalent to the geodesic flows on the L^2 sphere (see Section 2) and since the natural geometry on L^2 is equivalent to the nonparametric Fisher geometry, it is worth asking whether these inference schemes are adapted to the nonparametric Fisher geometry in a similar way to Riemannian HMC's adaptation to the parametric Fisher geometry.

Indeed this is the case, and it is a simple consequence of Proposition 1 in the supplementary file and the isometric relationship between square-integrable functions and square-summable sequences induced by any orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$ with completion L^2. Denote the space of square-summable sequences and its sphere

$$\ell^2 = \left\{ q = \{q_i\}_{i=1}^{\infty} \big| \, \langle q, q \rangle_{\ell^2} = \sum_{i=1}^{\infty} q_i^2 < \infty \right\},$$
$$\mathcal{S}^{\infty} = \left\{ q \in \ell^2 \big| \, \langle q, q \rangle_{\ell^2} = \sum_{i=1}^{\infty} q_i^2 = 1 \right\}.$$

Then it follows from the orthonormality of $\{\phi_i\}_{i=1}^{\infty}$ that $(L^2, \langle \cdot, \cdot \rangle_{L^2}) \cong (\ell^2, \langle \cdot, \cdot \rangle_{\ell^2})$, since for any arbitrary function $q = q(\cdot) \in L^2$,

$$\langle q, q \rangle_{L^2} = \int q(x)^2 \mu(dx) = \int \big(\sum_{i=1}^{\infty} q_i \phi_i(x) \big)^2 \mu(dx)$$
$$= \sum_{i=1}^{\infty} q_i^2 = \langle q, q \rangle_{\ell^2} \, .$$

It is an immediate result that the respective spheres are also isometric, i.e. $(\mathcal{Q}, \langle \cdot, \cdot \rangle_{L^2}) \cong (\mathcal{S}^{\infty}, \langle \cdot, \cdot \rangle_{\ell^2})$, and hence, by Proposition 1, the following result holds.

Lemma 1. *Given an orthonormal basis for L^2, the space of density functions equipped with the Fisher metric is isometric to the sphere $\mathcal{S}^{\infty}$ with its natural Euclidean metric, i.e. $(\mathcal{P}, g_F(\cdot, \cdot)) \cong (\mathcal{S}^{\infty}, \langle \cdot, \cdot \rangle_{\ell^2})$. (See the supplementary file for the proof.)*

Our goal is to show that spherical HMC can be adapted to the nonparametric Fisher geometry in the infinite-dimensional limit. Given that the geodesic paths followed by spherical HMC converge to geodesics on $\mathcal{S}^{\infty}$, Lemma 35 will imply that these paths correspond to geodesics on $(\mathcal{P}, g_F(\cdot, \cdot))$.

Lemma 2. *Geodesic flows on the finite sphere $\mathcal{S}^{I-1}$ converge to geodesic flows on the infinite-dimensional sphere $\mathcal{S}^{\infty}$ as $I \to \infty$. (See the supplementary file for the proof.)*

We are now ready to connect Riemannian HMC and spherical HMC in the infinite-dimensional limit (where the latter is applied to the square-root density estimation

problem). To make this relationship as clear as possible, we introduce a different (but equivalent) definition of a geodesic based on the calculus of variations (in contrast to the null acceleration definition from Lemma 1). Assume that two points A and B are close together in a small open set of Riemannian manifold $(\mathcal{M}, g(\cdot,\cdot))$. Let $\Gamma : [a,b] \times (-\epsilon,\epsilon) \to \mathcal{M}$ be a family of curves $\gamma_s : [a,b] \to \mathcal{M}$ satisfying $\gamma_s(a) = A$ and $\gamma_s(b) = B$ for all $s \in (-\epsilon,\epsilon)$. Then γ is a geodesic if it minimizes the energy functional

$$E(\gamma) = \frac{1}{2}\int_a^b g_{\gamma(t)}\big(\dot\gamma(t),\dot\gamma(t)\big)\,dt\,,$$

and thus satisfies $\frac{d}{ds}E(\gamma_s) = 0$.

For a parametric family of distributions $\mathcal{P}_\theta$ equipped with the Fisher metric, the *parametric Fisher energy* takes the form

$$\begin{aligned} E(\theta) =& \frac{1}{2}\int_a^b g_{\theta(t)}\big(\dot\theta(t),\dot\theta(t)\big)_F\,dt \\ =& \frac{1}{2}\int_a^b \nabla_\theta \ell(\theta(t))^T\,\mathcal{I}(\theta(t))^{-1}\,\nabla_\theta \ell(\theta(t))\,dt\,, \end{aligned}$$

where $\mathcal{I}(\theta)$ is the Fisher information, and $\ell(\theta) = \log p(\theta)$. On the other hand by Lemmas 1 and 5, the *nonparametric Fisher energy* for a family of curves in $\mathcal{P}$ takes the form

$$\begin{aligned} E(p) =& \frac{1}{2}\int_a^b g_{p(t)}\big(\dot p(t),\dot p(t)\big)_F\,dt \\ =& \frac{1}{2}\int_a^b \langle \dot q(t),\dot q(t)\rangle_{L^2}\,dt \\ =& \frac{1}{2}\int_a^b \langle \dot q(t),\dot q(t)\rangle_{\ell^2}\,dt \end{aligned}$$

where $q = \sqrt{p} = \sum_{i=1}^\infty q_i\phi_i(\cdot)$.

Theorem 1. *Let $q(\cdot) = \sqrt{p(\cdot)} \in \mathcal{Q}$ be a square-root density function with expansion satisfying*

$$\begin{aligned} q(\cdot) =& \sum_{i=1}^\infty q_i\phi_i(\cdot)\,, \textit{and} \\ 1 =& \int_{\mathcal{D}} q(x)^2\,\mu(dx) = \sum_{i=1}^\infty q_i^2\,, \end{aligned}$$

with random, real-valued coefficients $q_i, i = 1,\ldots,\infty$. Then, in the infinite-dimensional limit, spherical HMC follows the nonparametric Fisher metric's geodesic flows in the same way that Riemannian HMC follows the Fisher metric's geodesic flows over the parametric family of distributions $\mathcal{P}_\theta$. (See the supplementary file for the proof).

5 RELATIONSHIP TO THE COX PROCESS

The χ^2-process density prior may be used to model the intensity function of a Cox process (Cox, 1955). The Cox process is a point process over a given domain such that each realization at point t is drawn from a Poisson distribution with intensity $\mu(s)$, where intensity function $\mu(\cdot)$ is itself a random process over the same given domain. Cox processes are useful for the analysis of spatial and time series data. Given $\mu(\cdot)$, the likelihood of such data $\{s_n\}_{n=1}^N$ is given by

$$p\big(\{s_n\}_{n=1}^N|\mu(\cdot)\big) = \exp\Big(-\int_{\mathcal{D}}\mu(s)\,ds\Big)\times\prod_{n=1}^N\mu(s_n)\,. \tag{5}$$

Bayesian inference on $\mu(\cdot)$ requires the calculation of two integrals, that over the parameter space and that from Equation (5). We make the latter integral trivial by modeling the intensity function as the product of a density function and a positively constrained random variable:

$$\mu(s) = M\times p(s) = M\times q(s)^2\,.$$

In this case, the likelihood $p\big(\{s_n\}_{n=1}^N|\mu(\cdot)\big)$ may be written as

$$\exp\Big(-\int_{\mathcal{D}} Mq(s)^2\,ds\Big)\times\prod_{n=1}^N Mq(s_n)^2,$$

which is equal to

$$\exp(-M)\,M^N\prod_{n=1}^N q(s_n)^2\,.$$

Since the likelihood factors in M and $q(\cdot)$, it follows that the two random variables will be independent in posterior distribution *if* they are specified to be independent in prior distribution. Indeed, M may even be given a conjugate prior: it is easy to see that

$$M\sim\Gamma(a,b)\,,\quad \text{implies}\quad M|N\sim\Gamma(a+N,b+1)\,.$$

Sampling from the joint posterior of $\mu(\cdot)$ is as simple as independently sampling M from its posterior and $q^2(\cdot)$ from the χ^2-process density sampler and then multiplying the two together. Such a model should be used with care. As a function of the data, the posterior distribution of M solely depends on N, which is itself a single realization from a Poisson distribution. Thus, our χ^2-process density prior for the Cox process is useful in situations where ample prior information on M is available.

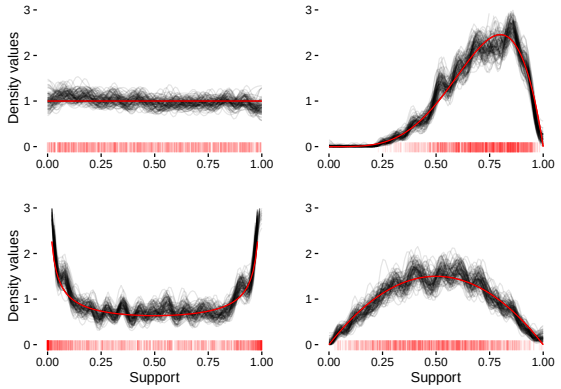

Figure 1: Each plot shows 100 posterior draws from the χ^2-process density sampler. 1,000 data samples were drawn from a different beta distribution for each plot. The generating pdf is given in red, and the red hash marks describe the actual data produced.

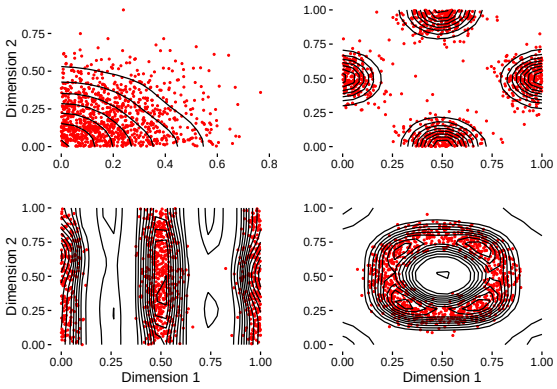

Figure 2: The contours (black) of the posterior median from 1,000 draws of the χ^2-process density sampler. Each posterior is conditioned on 1,000 data points (red).

6 EMPIRICAL RESULTS

Here we apply the χ^2-process density model to both simulated and real-world data. As stated in Section 3.1, the eigen-pairs corresponding to the GP with covariance operator (4) depend on both the dimension and the area of $\mathcal{D}$. When $\mathcal{D}$ is the one-dimensional unit interval, the eigen-pairs are given by

$$\lambda_i^2 = \sigma^2(\alpha + \pi^2 i^2)^{-s}, \quad \text{and} \quad \phi_i(x) = \sqrt{2}\cos(\pi\, i\, x),$$

for $i \geq 0$. For the two-dimensional unit square $\mathcal{D} = [0,1] \times [0,1]$, the eigen-pairs are given by

$$\begin{aligned} \lambda_i^2 = &\quad \sigma^2\big(\alpha + \pi^2(i_1^2 + i_2^2)\big)^{-s}, \\ \phi_i(x) = &\quad 2\cos(\pi\, i_1\, x_1)\cos(\pi\, i_2\, x_2), \end{aligned}$$

for $i_1, i_2 \geq 0$, where i_1 and i_2 are indices for the first and second dimensions of the domain respectively. See Beskos et al. (2016) for a similar approach.

6.1 SIMULATED EXPERIMENTS

Figure 1 depicts 1,000 data points (red hash marks) drawn from four different beta distributions (red) along with 100 MCMC draws from the posterior distribution based on the χ^2-process density model. From left to right and top to bottom, the beta distribution parameters are $(1,1)$, $(5,2)$, $(.5,.5)$, and $(2,2)$. Note that while the individual posterior draws adhere closely to the sampled data, the variability in the posterior draws accounts for uncertainty and gives good coverage to the true density. The hyperparameter settings for the top-left plot is given by $(\sigma, \alpha, s) = (.5, 1, 1)$, and $(\sigma, \alpha, s) = (.5, .5, .8)$ is the hyperparameter setting for the rest. We set $I = 30$ for each example, and 10,000 thinned MCMC iterations were used to make each figure.

Figure 2 depicts 1,000 data points (red) drawn from four different distributions on the unit square along with the contours of the pointwise median of 1,000 posterior draws from the χ^2-process density model. The data in the first three plots was generated using truncated Gaussians and mixtures of truncated Gaussians. The data for the last plot were generated by Gaussian noise added to the uniform distribution on the circle. The model adapts easily to multimodal and patterned data samples. For all examples, the hyperparameters were fixed to $(\sigma, \alpha, s) = (.9, .1, 1.1)$, and $0 \leq i_1, i_2 \leq 5$ for each example.

6.2 REAL-WORLD EXPERIMENTS

Figure 3 features the British coal mine disaster data set, in which the dates of 191 disasters are recorded between the years of 1851 and 1967. In both plots, the dates are given in red. Two comparisons are implied by the figure. The first is a comparison between the variability of 100 posterior draws based on 191 data points (left plot) with the variability in 100 posterior draws based on 1,000 data points, as in Figure 1. One sees much less variability in the latter. The other comparison is between the close fit exhibited in the posterior draws of the left plot compared to the smooth fit shown by the pointwise quantiles (median, black; .25, blue; .75, blue). As we can see, our method is valid for modeling densities without periodic tendencies, despite the specific form of the basis. Both plots are based on 10,000 thinned MCMC iterations, with hyperparameter settings $(\sigma, \alpha, s) = (.5, .5, .8)$ with $I = 30$.

Figure 4 features Hutchings' bramble canes data (red) (Hutchings, 1978), consisting of the locations of 823

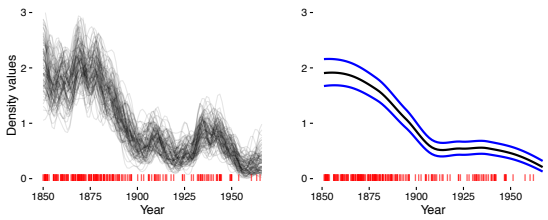

Figure 3: Coal mining disasters data: the left figure shows 100 posterior draws from the χ^2-process density model (gray) over 191 vertical lines (red) marking the precise date of each disaster. The right figure shows the pointwise median (black) for the same sample as well as pointwise quantile bands (blue).

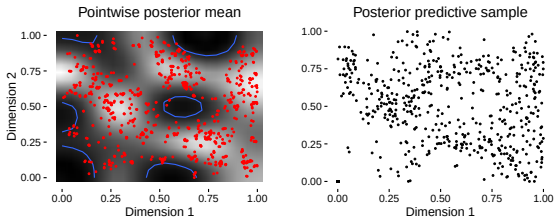

Figure 4: Hutchings' bramble canes data: the first figure depicts the 823 bramble canes (red), a heatmap of the pointwise posterior mean (black is low, white is high), and a single contour at density 0.3 (blue) including all but a few points. The second figure shows 823 draws from the χ^2-process density posterior predictive distribution.

bramble canes in a square plot. The left figure contains a heatmap of the pointwise posterior mean of the χ^2-process density model, where black pertains to low density and white pertains to high density. Finally, a single contour (blue) at density level 0.3 divides the majority of points from areas of extremely low density. The hyperparameters were set to $(\sigma, \alpha, s) = (2, .01, 1.1)$ with $0 \leq i_1, i_2 \leq 5$, and the posterior sample featured 10,000 MCMC iterations. The right figure features 823 draws from the posterior predictive distribution of the χ^2 process density model. Each draw from the posterior predictive distribution was obtained by randomly selecting one posterior draw from the χ^2 process density model. Since this single posterior sample is itself a density function, one can then sample from its corresponding distribution using a rejection sampling scheme. There is a remarkable similarity between the posterior predictive sample (right, black) and the bramble canes data (left, red).

7 DISCUSSION

We have presented a nonparametric extension to the parametric Fisher geometry and showed that this generalization is consistent with its parametric predecessor. To do so, the set of probability density functions over a given domain was defined to be an infinite-dimensional smooth manifold where each point is itself a density function. This manifold becomes a Riemannian manifold when equipped with the nonparametric Fisher information metric and is then identified with the infinite-dimensional sphere. We demonstrated one application of this approach in the form of Bayesian nonparametric density estimation. The resulting χ^2-process density model is flexible and computationally efficient: it is amenable to HMC and, in comparison to the cubic scaling of GP competitors, scales linearly in the number of data points. Of course, there is nothing *a priori* restricting the prior to be Gaussian. Also, an important next step is placing a prior on the number of basis functions to use, as is done in (Cotter et al., 2013).

The theoretical and methodological results presented in this paper are merely first steps in exploiting the simple geometry implied by the nonparametric Fisher metric.

Acknowledgement

This work is supported by NSF grant DMS 1622490 and NIH grant R01 MH115697.

References

Gourieroux, Christian and Alain Monfort (1995). *Statistics and econometric models*. Vol. 1. Cambridge University Press.

Jeffreys, Harold (1946). "An invariant form for the prior probability in estimation problems". In: *Proceedings of the Royal Society of London a: mathematical, physical and engineering sciences*. Vol. 186. 1007. The Royal Society, pp. 453–461.

Le Cam, Lucien (2012). *Asymptotic methods in statistical decision theory*. Springer Science & Business Media.

Fisher, Ronald Aylmer (1925). "Theory of statistical estimation". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 22. 05. Cambridge Univ Press, pp. 700–725.

Rao, C Radhakrishna (1945). "Information and accuracy attainable in the estimation of statistical parameters". In: *Bull. Calcutta Math. Soc* 37.3, pp. 81–91.

Amari, Shun-ichi and Hiroshi Nagaoka (2007). *Methods of information geometry*. Vol. 191. American Mathematical Soc.

Srivastava, Anuj, Ian Jermyn, and Shantanu Joshi (2007). "Riemannian analysis of probability density functions with applications in vision". In: *Computer Vision and*

Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, pp. 1–8.

Chen, Tian, Jeffrey Streets, and Babak Shahbaba (2015). "A Geometric View of Posterior Approximation". In: *arXiv preprint arXiv:1510.00861*.

Itoh, Mitsuhiro and Hiroyasu Satoh (2015). "Geometry of Fisher information metric and the barycenter map". In: *Entropy* 17.4, pp. 1814–1849.

Kurtek, Sebastian and Karthik Bharath (2015). "Bayesian sensitivity analysis with the Fisher–Rao metric". In: *Biometrika* 102.3, pp. 601–616.

Srivastava, Anuj and Eric P Klassen (2016). *Functional and shape data analysis*. Springer.

Peter, Adrian M, Anand Rangarajan, and Mark Moyou (2017). "The Geometry of Orthogonal-Series, Square-Root Density Estimators: Applications in Computer Vision and Model Selection". In: *Computational Information Geometry*. Springer, pp. 175–215.

Pinheiro, Aluisio and Brani Vidakovic (1997). "Estimating the square root of a density via compactly supported wavelets". In: *Computational Statistics & Data Analysis* 25.4, pp. 399–415.

Müller, Peter and Brani Vidakovic (1998). "Bayesian inference with wavelets: Density estimation". In: *Journal of Computational and Graphical Statistics* 7.4, pp. 456–468.

Hong, Xia and Junbin Gao (2016). "A Fast Algorithm to Estimate the Square Root of Probability Density Function". In: *Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV 33*. Springer, pp. 165–176.

Williams, Christopher KI and Carl Edward Rasmussen (1996). "Gaussian processes for regression". In: *Advances in neural information processing systems*, pp. 514–520.

Wang, Limin (2008). "Karhunen-Loeve expansions and their applications." PhD thesis. London School of Economics and Political Science (United Kingdom).

Dashti, Masoumeh and Andrew M Stuart (2013). "The Bayesian approach to inverse problems". In: *arXiv preprint arXiv:1302.6989*.

Cotter, Simon L, Gareth O Roberts, Andrew M Stuart, David White, et al. (2013). "MCMC methods for functions: modifying old algorithms to make them faster". In: *Statistical Science* 28.3, pp. 424–446.

Beskos, Alexandros, Mark Girolami, Shiwei Lan, Patrick E Farrell, and Andrew M Stuart (2016). "Geometric MCMC for infinite-dimensional inverse problems". In: *Journal of Computational Physics*.

Lan, Shiwei, Bo Zhou, and Babak Shahbaba (2014). "Spherical Hamiltonian Monte Carlo for constrained target distributions". In: *JMLR workshop and conference proceedings*. Vol. 32. NIH Public Access, p. 629.

Murray, Iain, David MacKay, and Ryan P Adams (2009). "The Gaussian process density sampler". In: *Advances in Neural Information Processing Systems*, pp. 9–16.

Rabier, Charles-Elie and Alan Genz (2014). "The supremum of Chi-Square processes". In: *Methodology and Computing in Applied Probability* 16.3, pp. 715–729.

Gunter, Tom, Michael A Osborne, Roman Garnett, Philipp Hennig, and Stephen J Roberts (2014). "Sampling for inference in probabilistic models with fast Bayesian quadrature". In: *Advances in neural information processing systems*, pp. 2789–2797.

Antoniak, Charles E (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". In: *The annals of statistics*, pp. 1152–1174.

Neal, Radford M (2000). "Markov chain sampling methods for Dirichlet process mixture models". In: *Journal of computational and graphical statistics* 9.2, pp. 249–265.

Byrne, Simon and Mark Girolami (2013). "Geodesic Monte Carlo on embedded manifolds". In: *Scandinavian Journal of Statistics* 40.4, pp. 825–845.

Cox, David R (1955). "Some statistical methods connected with series of events". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 129–164.

Longford, Nicholas (1987). "A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects". In: *ETS Research Report Series* 1987.1.

Girolami, Mark and Ben Calderhead (2011). "Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.2, pp. 123–214.

Efron, Bradley (1978). "The geometry of exponential families". In: *The Annals of Statistics* 6.2, pp. 362–376.

Chung, Kai Lai (2013). *Lectures from Markov processes to Brownian motion*. Vol. 249. Springer Science & Business Media.

Dryden, Ian L et al. (2005). "Statistical analysis on high-dimensional spheres and shape spaces". In: *The Annals of Statistics* 33.4, pp. 1643–1665.

Shahbaba, Babak, Shiwei Lan, Wesley O Johnson, and Radford M Neal (2014). "Split Hamiltonian Monte Carlo". In: *Statistics and Computing* 24.3, pp. 339–349.

Holbrook, Andrew J., Philippe Lemey, et al. (2020). "Massive Parallelization Boosts Big Bayesian Multidimensional Scaling". In: *Journal of Computational and Graphical Statistics* 0.0, pp. 1–14.

Hutchings, Michael J (1978). "Standing crop and pattern in pure stands of Mercurialis perennis and Rubus fruticosus in mixed deciduous woodland". In: *Oikos*, pp. 351–357.