Statistica Sinica Preprint No: SS-2018-0231	
Title	A Two-Step Geometric Framework For Density Modeling
Manuscript ID	SS-2018-0231
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0231
Complete List of Authors	Sutanoy Dasgupta Debdeep Pati and Anuj Srivastava
Corresponding Author	Sutanoy Dasgupta
E-mail	s.dasgupta@stat.fsu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

Statistica Sinica

A Two-Step Geometric Framework For Density Modeling

Sutanoy Dasgupta, Debdeep Pati, and Anuj Srivastava

Florida State University and Texas A& M University

Abstract: We introduce a novel two-step approach for estimating a probability density function (*pdf*), given its samples, with the second and important step coming from a geometric formulation. The procedure obtains an arbitrary initial estimate which it transforms using a warping function to reach the final estimate. The initial estimate is intended to be computationally fast, albeit suboptimal; however, but its warping creates a larger, flexible class of density functions, resulting in a substantially improved estimate. The optimal warping is determined by mapping warping functions to the tangent space of a Hilbert sphere, which is a vector space with elements that can be expressed using an orthogonal basis. Using a truncated basis expansion, we estimate the optimal warping under a (penalized) likelihood criterion and obtain the final density estimate. This framework is introduced for univariate unconditional *pdf* estimations, and then extended to include conditional *pdf* estimations. The approach avoids many of the computational pitfalls associated with classical conditional-density estimation methods, without sacrificing estimation performance. We derive the asymptotic convergence rates of our density estimator, and demonstrate this approach using synthetic data sets and real data, on the relation between a toxic metabolite on pre-term birth.

Key words and phrases: conditional density; density estimation; warped density; Hilbert

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

sphere; sieve estimation; tangent space; weighted likelihood maximization

1. Introduction

The estimation of probability density functions (*pdfs*) is an important and well-studied field of research in statistics. The most basic problem in this area is that of a univariate *pdf* estimation from independent and identically distributed *iid* samples, henceforth referred to as an unconditional density estimation. Another important problem is that of a conditional density estimation, where we need to characterize the behavior of the response variable for different values of the predictors.

Owing to the importance of *pdf* estimations in statistics and related disciplines, numerous solutions have been proposed for each of these problems. While the earliest works focused on parametric solutions, the trend over the last three decades has been to use a nonparametric approach, because it minimizes making assumptions about the underlying density (and about the relationships between the variables for conditional and joint densities). The most common nonparametric techniques are kernel based; refer to Rosenblatt (1956), Hall et al. (1991), Sheather & Jones (1991) and Li & Racine (2007) for further information. Related to these approaches are the "tilting" or "data sharpening" techniques for unconditional density estimations; see, for example, Hjort & Glad (1995), Doosti

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

& Hall (2016), and the references therein. Kernel methods are particularly powerful in a univariate setting. However, as the number of variables increases, these methods tend to become computationally inefficient owing to the complexities of bandwidth selection, especially in the case of a conditional density estimation.

1.1 Two-Step Approaches for Density Estimations

Another common approach used for *pdf* estimation, and the one employed in this study, is the two-step estimation procedure discussed in Leonard (1978), Lenk (1988, 1991), Tokdar, Zhu & Ghosh (2010), and Tokdar (2007). The first step estimates an initial *pdf*, say f_p, from the data, perhaps restricting it to belonging to a parametric family. In the second step, we improve upon this estimate by deriving a function $w > 0$ that depends on the initial estimate f_p, and obtaining a final estimate using $w(x)f_p(x)/\int_y w(y)f_p(y)dy$. Thus, the second step involves estimating an optimal w in order to reach the overall estimate. In a Bayesian context, the function w is often assigned a Gaussian process prior. While this approach is quite comprehensive, the calculation of the normalization constant at every step makes the computation very cumbersome. The two-step procedures can also be adapted to estimate conditional density functions. Here we estimate the conditional mean function, and then estimate the conditional density of the residuals, as in Hansen (2004). More recently, Bayesian methods

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

based on mixture models and latent variables for estimating *pdfs* have received increased attention, primarily as a result of their excellent practical performance and an increasingly rich set of algorithmic tools for sampling the posterior using Markov chain Monte Carlo (MCMC) methods. See Escobar & West (1995), Müller, Erkanli & West (1996), MacEachern & Müller (1998), Kalli, Griffin & Walker (2011), Jain & Neal (2012), Kundu & Dunson (2014) and Bhattacharya, Pati & Dunson (2010) among others. However, these methods also incur the very high computational cost typically associated with the MCMC algorithms. Applications of flexible Bayesian models for conditional densities are discussed in MacEachern (1999), De Iorio et al. (2004), Griffin & Steel (2006), Dunson, Pillai & Park (2007), Chung & Dunson (2009) and Norets & Pelenis (2012), among others. Although the literature suggests that such methods based on mixture models have several attractive properties, they lack interpretability. Furthermore, the MCMC solutions for model fitting tend to be overly complicated and expensive.

1.2 A Geometric Two-Step Approach

In this study, we pursue a geometric two-step approach that is applicable to both conditional and unconditional density estimations. Our main motivation is to develop an efficient estimation procedure that attains good estimation perfor-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

mance. This approach differs from the previously described two-step procedure in that the transformation of f_p (in the second step) is now based on the action of a diffeomorphism group, as follows. Let f_p be a strictly positive univariate density on the interval $[0, 1]$; here f_p serves as an initial estimate of the *pdf*. Let Γ be the set of all positive diffeomorphisms from $[0, 1]$ to itself, that is, $\Gamma = \{\gamma | \gamma \text{ is differentiable}, \gamma^{-1} \text{ is differentiable}, \dot{\gamma} > 0, \gamma(0) = 0, \gamma(1) = 1\}$. The elements of Γ play the role of warping functions, or transformations of f_p. Given $\gamma \in \Gamma$, the transformation of f_p is defined by $(f_p * \gamma) = (f_p \circ \gamma)\dot{\gamma}$. Henceforth, we refer to this transformation as the *warping* of f_p, and to the resulting *pdf* f as a *warped density*. This mapping is comprehensive, in the sense that we can change from any positive *pdf* to any other positive *pdf* using an appropriate γ. Note that because $\int_0^1 f_p(\gamma(x))\dot{\gamma}(x)dx = 1$, there is no need to normalize this transformation. However, the difficulty of estimating the normalizing constant now shifts to the problem of estimating over Γ, which poses some challenges because Γ is a nonlinear manifold. Note that diffeomorphisms as transformations of a *pdf* have been used in the past, albeit with a different setup and scope; see, for example, Saoudi, Hillion & Ghorbel (1994), and Saoudi, Ghorbel & Hillion (1997). In addition, the notion of a transformation between *pdf*s has been used in the literature on *optimal transport,* as in Tabak & Turner (2013) and Tabak & Trigila (2014). However, in this case, the transport is achieved using an iterated

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

composition of maps, and not through an optimization over Γ, as we do here. There are two parts to this.

1. **Univariate PDF Estimation**: We start with a framework for estimating an unconditional univariate *pdf* defined on $[0, 1]$. This unit interval setting helps explain and illustrate the main components of the framework. In addition, the proposed geometric framework is naturally univariate, in the sense that the transformation defined earlier acts on univariate density shapes, making it a logical starting point. In this simple setup, the approach delivers excellent performance, while avoiding a heavy computational cost, that is comparable with that of the standard kernel methods, even at very low sample sizes. The framework is then extended to univariate densities with unknown support by scaling the observation domain to $[0, 1]$. A defining characteristic of this warping transformation is that the initial estimate can be constructed in any way, whether parametric (e.g. Gaussian) or nonparametric (e.g., kernel estimate), and is allowed to be a suboptimal estimate of the true density.

2. **Conditional Density Estimation**: The second part of the paper extends the framework to the estimation of the conditional density $f(y|x)$ from $\{(y_i, x_i) : i = 1, \ldots, n, y \in \mathbb{R}, x \in \mathbb{R}^d, d \geq 1\}$. Here we start with a nonparametric mean regression model of the form $y_i = m(x_i) + \epsilon_i$, where

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and $m(\cdot)$ is estimated using a standard nonparametric estimator to obtain an initial conditional density estimate $f_{p,x} \equiv \mathcal{N}(\hat{m}(x), \hat{\sigma}^2)$ at location x. Then, $f_{p,x}$ is warped using the warping function γ_x into a final conditional density estimate. Naturally, the choice of $\gamma_x \in \Gamma$ varies with the predictor x. The selection of γ_x is based on a weighted-likelihood objective function that borrows information from the neighborhood of the location x at which the conditional density is being evaluated.

The main contributions of this paper as as follows.

1. **Transformation-Based Estimation**: We introduce a two-step density estimation framework based on the group action of Γ on the space of densities.

2. **Geometry of** Γ: The framework uses the differential geometry of Γ to map its elements to a subset of a Hilbert space, allowing for a basis expansion and the application of standard optimization tools for estimating warping functions.

3. **Conditional Density Estimation**: We derive an efficient framework for estimating conditional densities, that delivers competitive practical performance and an improved computational cost, compared with those of the standard kernel techniques.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

The rest of the paper is organized as follows. Section 2 outlines the general framework for a univariate unconditional density estimation, and Section 3 presents an asymptotic analysis of this estimator. Section 4 presents simulation results to illustrate the framework. Section 5 develops the theory for the conditional density estimation, and illustrates the properties of the proposed method using simulated data sets and real data.

2. Proposed Framework

In this section, we develop a two-step framework for estimating a univariate unconditional *pdf*. First we introduce some notation. Let $\mathscr{F}$ be the set of all strictly positive univariate pdfs on $[0, 1]$. (Note that this framework can be extended easily to densities with unknown support; see the Supplementary Material Section 6.2.) Let $f_0 \in \mathscr{F}$ denote the underlying true density, and $X_i \sim f_0$, for $i = 1, 2, \ldots, n$, be independent samples from f_0. Furthermore, let $\mathscr{F}_p$ be a predetermined subset of $\mathscr{F}$, such that an optimal element (based on the likelihood, or any other desired criterion) on $\mathscr{F}_p$ is relatively easy to compute; call it f_p. For instance, any parametric family with a simple maximumlikelihood estimator is a good candidate for f_p. Similarly, kernel density estimates work well because they are computationally efficient and robust in univariate setups.

Next, we define a warping-based transformation of the elements of $\mathscr{F}_p$, us-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

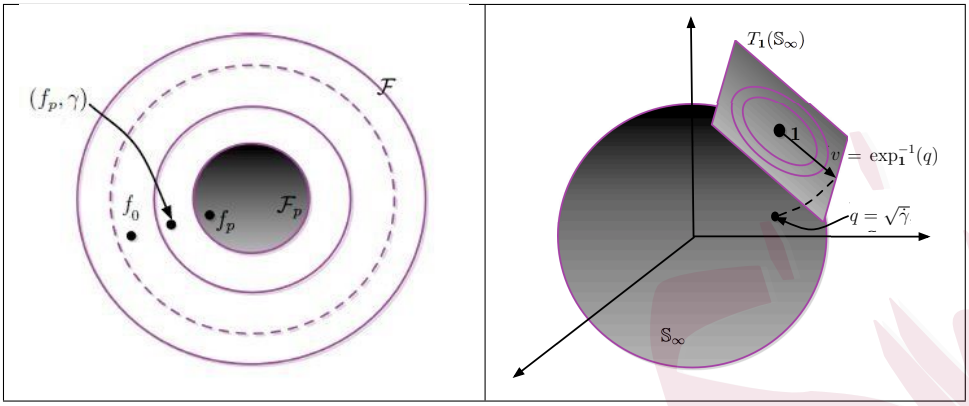

Figure 1: *Left: The true pdf f_0 is estimated by transforming an initial estimate f_p by the warping function γ. The larger the set of allowed γs, the better the estimate is. Right: Representing warping function γ as an element of the tangent space $T_{\mathbf{1}}(\mathbb{S}_\infty^+)$.*

ing the elements of Γ defined earlier. Note that Γ is an infinite-dimensional manifold that has a group structure under composition as the group operation. That is, for any $\gamma_1, \gamma_2 \in \Gamma$, the composition $\gamma_1 \circ \gamma_2 \in \Gamma$. The identity element of Γ is given by $\gamma_{\text{id}}(t) = t$, and for every $\gamma \in \Gamma$, there is a function $\gamma^{-1} \in \Gamma$, such that $\gamma \circ \gamma^{-1} = \gamma_{\text{id}}$. For any $f_p \in \mathscr{F}_p$ and $\gamma \in \Gamma$, define the mapping $* : \mathscr{F} \times \Gamma \to \mathscr{F} : (f_p * \gamma) = (f_p \circ \gamma)\dot{\gamma}$ as given earlier. This mapping is akin to the change of variable formula for densities. The importance of this mapping comes from the following result.

Proposition 1. *The mapping $* : \mathscr{F} \times \Gamma \to \mathscr{F}$, specified above, forms an action of Γ on $\mathscr{F}$. Furthermore, this action is transitive. In other words, one can reach any element of $\mathscr{F}$ from any other element of $\mathscr{F}$ using an appropriate element of*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

Γ.

Proof: We can verify the two properties in the definition of a group action: (1) For any $\gamma_1, \gamma_2 \in \Gamma$ and $f \in \mathscr{F}$, we have $((f * \gamma_1) * \gamma_2) = (((f \circ \gamma_1)\dot{\gamma}_1) \circ \gamma_2)\dot{\gamma}_2 = (f * \gamma_1 \circ \gamma_2)$. (2) For any $f \in \mathscr{F}$, $(f * \gamma_{\text{id}}) = f$. To show transitivity, we need to show that, given any $f_1, f_2 \in \mathscr{F}$, there exists a $\gamma \in \Gamma$, such that $(f_1 * \gamma) = f_2$. If F_1 and F_2 denote the cumulative distribution functions associated with f_1 and f_2, respectively, then the desired γ is simply $F_1^{-1} \circ F_2$. Because f_1 is strictly positive, F_1^{-1} is well defined, and γ is uniquely specified. Furthermore, because f_2 is strictly positive, we have $\dot{\gamma} > 0$ and $\gamma \in \Gamma$. □

This result implies that, together, the pair $(f_p * \gamma)$ spans the full set $\mathscr{F}$ if γ is chosen freely from Γ. However, if one uses a proper submanifold of Γ instead of the full Γ, we may not reach the desired f_0, but only approximate it in some way. This intuition is depicted pictorially in the left panel of Figure 1, where the inner disk denotes the set $\mathscr{F}_p$. The increasing rings around $\mathscr{F}_p$ represent the set $\{(f_p * \gamma) | f_p \in \mathscr{F}_p\}$, with γ belonging to progressively larger dimensional submanifolds of Γ. As the submanifolds approach the full space Γ, the corresponding approximation approaches f_0. The submanifolds are introduced formally in the next subsection. Additional information can be found in Section S1.1 of the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

2.1 Finite-Dimensional Representation of Warping Functions

Given an initial estimate f_p, we now determine an optimal γ, such that the warped density $(f_p \circ \gamma)\dot{\gamma}$ becomes the final estimate under the chosen criterion. However, solving an optimization problem over Γ faces two main challenges. First, Γ is a nonlinear manifold, and second, it is infinite-dimensional. Here, Γ is a nonlinear manifold because it is not a vector space. (That is, an arbitrary linear combination of elements of Γ is, typically, not in Γ.) We handle the nonlinearity by forming a map from Γ to a vector space, and the infinite dimensionality by selecting a finite-dimensional subspace of this vector space. Together, these two steps are equivalent to finding a family of finite-dimensional submanifolds of Γ that can be *flattened* into vector spaces. This enables us to represent a variable γ using elements of a Euclidean vector space and to apply standard optimization procedures. This representaion, explained in detail below, enjoys important advantages over direct approximations of γ; for further information, see the Supplementary Material Section 6.2.

To flatten Γ locally, we define a function $q_\gamma : [0,1] \to \mathbb{R}$, $q_\gamma(t) = \sqrt{\dot{\gamma}(t)}$, termed the *square root slope function* (SRSF) of $\gamma \in \Gamma$. (For a discussion on SRSFs of general functions, refer to Chapter 4 of Srivastava & Klassen (2016)). To understand the nature of this relation $\gamma \to q_\gamma$, consider the set $\mathcal{Q}_\gamma = \{q : [0,1] \to \mathbb{R} | \int_0^t q^2(s)ds = \gamma(t)\}$, consisting of all functions that can be mapped

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

back to the same γ. Clearly, $q_\gamma \in \mathcal{Q}_\gamma$ and, hence, the set is always nonempty. Secondly, for any pair $\gamma_1 \neq \gamma_2$, $\mathcal{Q}_{\gamma_1}$ and $\mathcal{Q}_{\gamma_2}$ are disjoint. We denote the unit Hilbert sphere by $\mathbb{S}_\infty \subset \mathbb{L}^2 = \{q : [0,1] \to \mathbb{R} | \int q^2(t)\, dt = 1\}$. Then, it is easy to see that, for all $\gamma \in \Gamma$, $\mathcal{Q}_\gamma \subset \mathbb{S}_\infty$. This is because for any $q \in Q_\gamma$, we have $\|q\|^2 = \int_0^1 q(t)^2 dt = \int_0^1 \dot{\gamma}(t) dt = \gamma(1) - \gamma(0) = 1$. The set $\mathbb{S}_\infty$ is a smooth manifold with known geometry under the $\mathbb{L}^2$ Riemannian metric (Lang (2012)). Although it is not a vector space, it can be easily flattened into a vector space (locally) owing to its constant curvature. A natural choice for flattening is the *retraction* to the vector space tangent to $\mathbb{S}_\infty$ at the point $\mathbf{1}$, which is a constant function with value 1. ($\mathbf{1}$ is the SRSF corresponding to $\gamma = \gamma_{\text{id}}(t) = t$.) The tangent space of $\mathbb{S}_\infty$ at $\mathbf{1}$ is an infinite-dimensional vector space, given by $T_{\mathbf{1}}(\mathbb{S}_\infty) = \{v \in \mathbb{L}^2([0,1], \mathbb{R}) | \int_0^1 v(t) dt = \langle v, \mathbf{1} \rangle = 0\}$. See the right panel of Figure 1 for an illustration of this idea. Next, we define the antipodal set of $\mathbf{1}$ on $\mathbb{S}_\infty$ to be the subset: $A_{\mathbf{1}} = \{q \in \mathbb{S}_\infty | \langle q, \mathbf{1} \rangle = -1\}$. Next, we define a bijective mapping between the set $\mathbb{S}_\infty / A_{\mathbf{1}}$ and the tangent space $T_{\mathbf{1}}(\mathbb{S}_\infty)$ using the inverse exponential map, defined as follows:

$$\exp_{\mathbf{1}}^{-1}(q) : \mathbb{S}_\infty / A_{\mathbf{1}} \longrightarrow T_{\mathbf{1}}(\mathbb{S}_\infty), \quad v = \exp_{\mathbf{1}}^{-1}(q) = \frac{\theta}{\sin(\theta)} (q - \mathbf{1}\cos(\theta)), \quad (2.1)$$

where $\theta = \cos^{-1}(\langle \mathbf{1}, q \rangle)$ is the arc-length from q to $\mathbf{1}$. The right panel of Figure

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

1 also shows the mapping from $\mathbb{S}_\infty/A_{\mathbf{1}}$ to $T_{\mathbf{1}}(\mathbb{S}_\infty)$. We impose a natural Hilbert structure on $T_{\mathbf{1}}(\mathbb{S}_\infty)$ using the standard inner product, $\langle v_1, v_2\rangle = \int_0^1 v_1(t)v_2(t)dt$. It is easy to check that because $\cos^{-1}(\langle \mathbf{1}, q\rangle) < \pi$, we have the norm $\|v\| = \sqrt{\int_0^1 v(t)^2 dt} = \theta < \pi$, for $v = \exp_{\mathbf{1}}^{-1}(q)$. Thus, the range of the inverse exponential map is not the entire $T_{\mathbf{1}}(\mathbb{S}_\infty)$, but a subset $T_{\mathbf{1}}^0(\mathbb{S}_\infty) = \{v \in T_{\mathbf{1}}(\mathbb{S}_\infty) : \|v\| < \pi\}$. In order to map points back from the tangent space to the unit Hilbert sphere, we use the exponential map given by:

$$\exp_{\mathbf{1}}(v) : T_{\mathbf{1}}^0(\mathbb{S}_\infty) \to \mathbb{S}_\infty, \quad \exp_{\mathbf{1}}(v) = \cos(\|v\|)\mathbf{1} + \frac{\sin(\|v\|)}{\|v\|}\, v. \tag{2.2}$$

Thus, for every $\gamma \in \Gamma$, there exists a set $V_\gamma = \exp_{\mathbf{1}}^{-1}(\mathcal{Q}_\gamma) \in T_{\mathbf{1}}^0(\mathbb{S}_\infty)$, such that $\exp_{\mathbf{1}}(V_\gamma) = \mathcal{Q}_\gamma$.

Finally, we can select any orthogonal basis $\mathcal{B} = \{b_j, j = 1, 2, \dots\}$ of the Hilbert space $T_{\mathbf{1}}^0(\mathbb{S}_\infty)$ to express its elements v by their corresponding coefficients; that is, $v(t) = \sum_{j=1}^\infty c_j b_j(t)$, where $c_j = \langle v, b_j\rangle$. The only restriction on the basis elements b_j is that they must be orthogonal to $\mathbf{1}$; that is, $\langle b_j, \mathbf{1}\rangle = 0$. For example, one can use the Fourier basis elements (excluding $\mathbf{1}$; of course). However, other bases, such as cosine basis, splines, and Legendre polynomials, can also be used. In the experimental studies, we use Meyer wavelets, which have attractive properties of infinite differentiability and support over all reals.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

Vermehren & de Oliveira (2015) provide a closed-form expression for Meyer wavelets and the scale function in the time domain, which enables us to use the basis set for the representation. However, Meyer wavelets are not naturally orthogonal to $\mathbf{1}$, and so need to be orthogonalized first; however, this can be done offline. Efromovich (2010) discusses different choices of basis functions, and advocates using trigonometric basis for functions with compact support. They also discuss how it is advantageous to always assume that the true density has a compact support and to scale it to the unit interval.

Given a basis set $\mathcal{B} = \{b_j, j = 1, 2, \dots\}$ and $\gamma \in \Gamma$, we can define an infinite-dimensional set $\mathcal{C}_\gamma = \{c = (c_1, c_2, \cdots) | \sum_{j=1}^{\infty} c_j b_j(t) \in V_\gamma\}$. However, we can use a truncated basis expansion to approximate the elements of the set $T_{\mathbf{1}}^0(\mathbb{S}_\infty)$ using finitely many coefficients. Suppose we use J basis elements to approximate the tangent space elements. Then, the approximating space of coefficients can be denoted by $\mathcal{C}^J = \{c \in \mathbb{R}^J | \sum_{j=1}^{J} c_j b_j(t) \in T_{\mathbf{1}}^0(\mathbb{S}_\infty)\}$. Note that $\mathcal{C}^J$ is a proper subset of $\mathbb{R}^J$ because it only contains elements satisfying $\| \sum_{j=1}^{J} c_j b_j(t) \| < \pi$. Using these two steps, we specify a finite-dimensional and, therefore, approximate representation of warpings. We define a composite

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

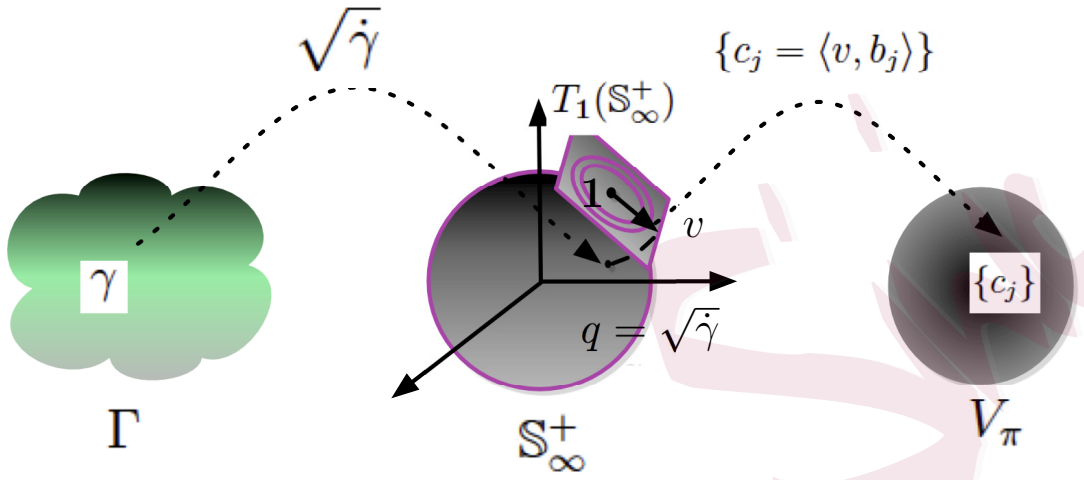

Figure 2: A graphic representation of Eqn. 2.3, leading to a map from V_π^J to Γ.

map $H : \mathcal{C}^J \to \Gamma$, as

$$\{c_j\} \in \mathcal{C}^J \xrightarrow{\{b_j\}} v = \sum_{j=1}^{J} c_j b_j \in T_{\mathbf{1}}^0(\mathbb{S}_\infty) \xrightarrow{\exp_{\mathbf{1}}} q \in \mathbb{S}_\infty \to \gamma(t) = \int_0^t q(s)^2 ds \,. \tag{2.3}$$

For any $c \in \mathcal{C}^J$, let γ_c denote the diffeomorphism $H(c)$. For any fixed J, the set $H(\mathcal{C}^J)$ forms a J-dimensional submanifold of Γ, henceforth denoted by Γ^J, on which we pose the estimation problem. As J goes to infinity, this submanifold Γ^J converges to the full group Γ.

With this setting, we can rewrite the estimation of the unknown density f_0,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

given an initial estimate f_p, as $\hat{f}(t) = f_p(\gamma_{\hat{c}}(t))\dot{\gamma}_{\hat{c}}(t), t \in [0,1]$, where $\gamma_{\hat{c}} = H(\hat{c})$ and

$$\hat{c} = \underset{c \in \mathcal{C}^J}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \left[\log \left(f_p \left(\gamma_c(X_i) \right) \dot{\gamma}_c(X_i) \right) \right] \right) . \tag{2.4}$$

This optimization problem is nonconvex. We use the standard MATLAB function `fminsearch` for the optimization; for a discussion of the algorithm, see *https://www.mathworks.com/help/optim/ug/fminsearch-algorithm.html*. The truncated basis approximation takes place in the tangent space representation of Γ, rather than in the original density space as is the case in Birgé & Massart (1998), Donoho et al. (1996), and several others.

3. Asymptotic Analysis and Convergence Rate Bounds

In this section, we determine the asymptotic convergence rate of the (maximum likelihood) density estimate $\hat{f}$, described according to (2.4) in Section 2.1, to the true underlying density f_0, using the sieve MLE proposed by Wong & Shen (1995). Let $\mathscr{F}$ denote the space of all univariate, strictly positive *pdf*s on $[0,1]$, as before. Let f_p be the initial density estimate obtained in the first step of the estimation process.

- Assumption 1: $f_0 : [0,1] \to \mathbb{R}^+$ is continuous and strictly positive.
- Assumption 2: f_0 belongs to either a Hölder or a Sobolev space of order

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

β.

- Assumption 3: $f_p : [0,1] \to \mathbb{R}^+$ is strictly positive and is Lipschitz continuous.

Note that in order to represent the entire space $\mathscr{F}$, we need a Hilbert basis with infinitely many elements. However, in practice, we use only a finite number of basis elements. Hence, we are actually optimizing over a subset of the space of density functions based on finitely many basis elements, and using this to approximate the true density. This subset is called the *approximating space*. Let n be the number of available observations. Let $\mathscr{F}_n$ be the approximating space of $\mathscr{F}$ when using $J = k_n$ basis elements for the tangent space $T_{\mathbf{1}}(\mathbb{S}^+_\infty)$, where k_n is some function of n. Let $f_p \in \mathscr{F}_p \subset \mathscr{F}$ denote the initial estimate satisfying Assumption 3. Examples of such f_p include Gaussian densities truncated to $[0,1]$, kernel density estimates with a plug-in bandwidth, and so on. Let $\mathscr{F}_n = \{f_p(\gamma)\dot{\gamma}, \gamma = H(c))|\ \ c \in \mathcal{C}^J \subset \mathbb{R}^{k_n}\}$, where H and $\mathcal{C}^J$ are defined in Section 2.1. As $n \to \infty, k_n \to \infty$ and, hence, $\mathscr{F}_n \to \mathscr{F}$. Let η_n be a sequence of positive numbers converging to zero. Let Z_i be the n observed data points scaled to the unit interval. We call an estimator $\hat{f} : [0,1] \to \mathscr{F}_n$ an η_n sieve MLE if

$$\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}(Z_i) \geq \sup_{p \in \mathscr{F}_n} \frac{1}{n}\sum_{i=1}^{n} \log p(Z_i) - \eta_n,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

In the proposed method, $\hat{f} : [0,1] \to \mathscr{F}_n$, as defined in (2.4), satisfies that $\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}(Z_i)$ is exactly $\sup_{p \in \mathscr{F}_n} \frac{1}{n}\sum_{i=1}^{n} \log p(Z_i)$. Therefore, $\hat{f}$ is a sieve MLE with $\eta_n \equiv 0$. Let $\|\cdot\|_r$ denote the $\mathbb{L}^r$ norm between functions. Then, the following theorem provides the asymptotic convergence rate for the sieve MLE $\hat{f}$.

Theorem 1. *Let $\epsilon_n^* = M_1 n^{-\beta/(2\beta+1)}\sqrt{\log n}$, for some constant M_1. If f_0 satisfies Assumptions 1 and 2, and f_p satisfies Assumption 3, then there exist constants C_1 and C_2, such that*

$$P(\|\hat{f}^{1/2} - f_0^{1/2}\|_2 \geq \epsilon_n^*) \leq 5\exp\big(-C_2 n(\epsilon_n^*)^2\big) + \exp\big(-\frac{1}{4}C_1 n(\epsilon_n^*)^2\big). \quad (3.1)$$

The proof of this theorem is deferred to the Supplementary Material. The idea of the proof hinges on proving the equivalence of the density space and the parameter space. That is, we show that if the estimated parameter is "close" to the true parameter corresponding to the true density, in some sense, then the corresponding estimated density is also "close" to the true density. The statement is stated formally and proved in Lemma S.1 in the Supplementary Material.

Note that the convergence rate presented in Theorem 1 is independent of the initial step f_p (up to constant terms) because the estimation problem is shifted to Γ, given a fixed choice of f_p. Intuitively, different initial choices of f_p will

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

result in different warping functions that would reshape f_p to the correct shape of f_0. Hence, the notion of a "true" warping function is identifiable only after specifying a fixed choice of f_p. However, once some f_p satisfying Assumption 3 is fixed, the convergence rate of $\hat{\gamma}$ to the "true" warping function and, hence, the convergence rate of $\hat{f}$ to f_0, is independent of f_p.

4. Simulation Studies

Next, we present the results of experiments in which we apply the univariate unconditional density estimation procedure to two simulated data sets. The code used in the experiments presented here can be found at `https://github.com/Sutanoy/Density-estimation-1`. The computations described here are performed on an Intel(R) Core(TM) i7-3610QM CPU processor, and the computational times are reported for each experiment. First, we compare the average performance of the proposed solution with two standard techniques: (1) kernel density estimates, with bandwidth selected using the unbiased cross-validation method, henceforth referred to as *kernel(ucv)*; and (2) a standard Bayesian technique that uses the function *DPdensity* in the R package `DPPackage`. We focus on the average performance of the various techniques over 100 independent samples from the true density. We use the MATLAB function *ksdensity* to determine the initial estimate f_p for our approach; *ksdensity* uses the naive

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

thumb rule for bandwidth selection and is computationally very fast, albeit sub-optimal. We consider sample sizes of $n = 25, 100$, and 1000 to study the effect of n on the estimation performance and the computational cost. The performance is evaluated using multiple norms: the $\mathbb{L}^2$ and$\mathbb{L}^1$ norm and the $\mathbb{L}^\infty$ norm, averaged over the 100 samples.

We borrow the first example from Tokdar (2007) and Lenk (1991), where $f_0 \propto 0.75\text{exp}(\text{rate} = 3) + 0.25\mathcal{N}(0.75, 1/8^2)$, a mixture of exponential and normal densities, truncated to the interval $[0, 1]$. Table 1 summarizes estimation performance and computation cost for these methods at different sample sizes. The mean and standard deviation values have been scaled by 100 for clarity. We find that when $n = 25$, the *kernel(ucv)* method performs slightly better than our method. However, for larger sample sizes, the warping-based method performs better overall. The computational cost of the proposed method, while higher than that of *kernel(ucv)*, is much less than that of *DPdensity* for larger sample sizes. We used Meyer wavelets as the basis set for the tangent space representation, and used Algorithm 1 (see Section S2 of Supplementary Material for details) to obtain an optimal number of basis elements. We also examined the performance using the Fourier basis and found very similar results.

For the second example, we use Example 10 from Marron & Wand (1992), who uses the claw density $f_0 = \frac{1}{2}\mathcal{N}(0, 1) + \sum_{l=0}^{4} \frac{1}{10}\mathcal{N}(\frac{l}{2} - 1, (0.1)^2)$. As before,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

Table 1: *The performance of the mixture of exponential and normal densities.*

Method:		DPDensity			Kernel(ucv)			Our Estimate		
n	Norm	Mean	std.dev.	Time	Mean	std.dev	Time	Mean	std.dev	Time
25	$\mathbb{L}^1$	37.26	8.63	4 sec	33.51	11.97	< 1 sec	34.37	11.11	12 sec
	$\mathbb{L}^2$	5.05	0.9		4.5	1.44		4.43	1.42	
	$\mathbb{L}^\infty$	1.64	0.21		1.44	0.47		1.28	0.48	
100	$\mathbb{L}^1$	22.87	5.32	18 sec	21.9	5.54	< 1 sec	19.69	5.48	12 sec
	$\mathbb{L}^2$	3.47	0.58		3.14	0.57		2.77	0.69	
	$\mathbb{L}^\infty$	1.49	0.2		1.23	0.24		1.04	0.32	
1000	$\mathbb{L}^1$	10.79	2.05	225 sec	11.57	2.14	< 1 sec	10.40	1.70	12 sec
	$\mathbb{L}^2$	1.83	0.24		1.67	0.23		1.65	0.33	
	$\mathbb{L}^\infty$	1.18	0.2		0.88	0.22		0.87	0.22	

Table 2: *The performance of the claw density.*

Method:		DPDensity			Kernel(ucv)			Our Estimate		
n	Norm	Mean	std.dev.	Time	Mean	std.dev	Time	Mean	std.dev	Time
25	$\mathbb{L}^1$	39.15	6.29	4 sec	17.06	2.33	1 sec	18.28	3.3	105 sec
	$\mathbb{L}^2$	5.46	0.48		2.09	0.3		2.41	0.43	
	$\mathbb{L}^\infty$	1.2	0.05		0.5	0.14		0.64	0.17	
100	$\mathbb{L}^1$	28.39	4.55	26 sec	8.54	2.38	1 sec	9.06	2.6	85 sec
	$\mathbb{L}^2$	4.31	0.46		1.18	0.28		1.3	0.35	
	$\mathbb{L}^\infty$	1.08	0.09		0.34	0.08		0.42	0.13	
1000	$\mathbb{L}^1$	19.28	1.63	331 sec	2.4	0.38	1 sec	2.46	0.43	71 sec
	$\mathbb{L}^2$	3.16	0.15		0.38	0.06		0.4	0.08	
	$\mathbb{L}^\infty$	0.83	0.04		0.14	0.03		0.15	0.04	

we employ Algorithm 1 (see Section S2 of the Supplementary Material) to find the optimal number of tangent basis elements J based on the AIC, with $J \leq 40$.

Table 2 summarizes the performance, showing that at $n = 1000$, the three methods perform similarly, especially *kernel(ucv)* and the warped density estimate. In fact, the warped density estimate and *kernel(ucv)* perform similarly

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

even at low sample sizes, whereas *DPdensity* performs worse. These results were obtained using the Fourier basis, but the results for the Meyer basis were similar. Note that the computation cost is highest for $n = 25$ for our method, and actually decreases as n increases. This is because, for small n, there is less information; thus it takes more time for the objective function to converge.

We also study the effect of the choice of the initial shape on the overall performance of the estimator, and compare the boundary performance with that of a standard kernel estimate; see Section S5 of the Supplementary Material.

5. Extension to Conditional Density Estimation

The idea of using diffeomorphisms to warp an initial density estimate, while maximizing the likelihood, extends naturally to conditional density estimation. Consider the following setup. Let X be a fixed d-dimensional random variable, with a positive density on its support. Let $Y \sim f_0(m(X), \sigma^2(X))$, where f_0 is the unknown conditional density that changes smoothly with X; $m(X)$ is the unknown mean function, assumed to be differentiable; and $\sigma^2(X)$ is the unknown variance, which may or may not depend on X. Here, Y is assumed to have a univariate continuous distribution, with support on an unknown interval $[A, B]$. We observe the pairs (Y_i, X_i), for $i = 1, \ldots, n$, and are interested in recovering the conditional density f_0.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

In order to initialize the estimation, we assume a nonparametric mean regression model of the form $y_i = m(x_i) + \epsilon_i$, $\epsilon_i \sim f_p(0, \sigma^2)$, where $m(\cdot)$ is estimated using a standard local linear regression, f_p is an initial estimate for the conditional density of the response variable, and σ^2 is estimated using the sample standard deviation of the residuals $\{Y_i - \hat{m}(X_i)\}$. We used the truncated normal density as f_p in the experiments presented later, but other choices are equally valid. In addition, we can choose any cost-efficient conditional density estimate directly as the initial guess. As was the case in the unconditional *pdf* estimation, it is not required that the initial estimate has a mean function close to the true mean function, or that it takes any particular form. The only requirement is that the initial conditional density should be continuous and bounded away from zero, and that the density should vary smoothly with X, in the sense that if x_1 and x_2 are close to each other, then the conditional *pdf* of $(Y|X = x_1)$ should be close to the conditional *pdf* of $(Y|X = x_2)$, under the $\mathbb{L}^2$ or some related metric. Then, the warped density estimate, for a warping function γ and location x_0, is $f_{w,x_0}(y|X = x_0) = f_p(\gamma(y), \hat{m}(x_0), \hat{\sigma}^2)\dot{\gamma}(y)$. Let F_{p,x_0} be the initial estimate of the conditional distribution function of Y, given $X = x_0$, for some given value of the predictor x_0. If F_{t,x_0} is the true conditional distribution function of Y, given $X = x_0$, then the true γ at location x_0 is $\gamma_{x_0} = F^{-1}_{p,x_0} \circ F_{t,x_0}$. Setting $f_{p,x_0} \equiv f_p(\hat{m}(x_0), \hat{\sigma}^2)$, we estimate

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

the optimal γ using the following weighted maximum likelihood estimation: $\hat{\gamma}_{x_0} = \text{argmax}_{\gamma\in\Gamma}\left(\sum_{i=1}^{n}\log\left[(f_{p,x_0}(\gamma(y_i)|x_i)\dot{\gamma})W_{x_0,i}\right]\right)$, where $W_{x_0,i}$ is the localized weight associated with the ith observation, calculated as

$$W_{x_0,i} = \frac{\mathcal{N}(\|X_i - x_0\|_2/h(x_0); 0, 1)}{\sum_{j=1}^{n}\mathcal{N}(\|X_j - x_0\|_2/h(x_0); 0, 1)},$$

where $\mathcal{N}(\cdot; 0, 1)$ is the standard normal *pdf*, and $h(x_0)$ is the parameter that controls the relative weights associated with the observations. Note that although we have used a Gaussian kernel to define the weights, any kernel can be used. However, the weights defined in this way result in a higher bias because information is being borrowed from all observations. As in Bashtannyk & Hyndman (2001), we allow only a specified fraction of the observations X_i to have a positive weight. However, using too small a fraction will result in unstable estimates and poor practical performance, because the effective sample size will be too small. Hence, we advocate using the nearest 50% of the observations (nearest to the target location) for borrowing information, and then calculating the weights for this smaller sample as before.

The parameter $h(x_0)$ is akin to the bandwidth parameter associated with traditional kernel methods for density estimation. A very large value of $h(x_0)$ distributes approximately equal weight to all observations, whereas a very small

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

value considers only those observations in a small neighborhood around x_0. Because $h(x_0)$ is scalar, we avoid the tremendous computational cost associated with obtaining cross-validated bandwidths in each predictor dimension, when the predictor dimension is high. When the predictor is one-dimensional, the parameter $h(x_0)$ is chosen according to the location x_0 using the following two-step procedure:

1. Compute a standard kernel density estimate $\hat{K}$ of the predictor space, using a fixed bandwidth chosen according to any standard criterion. Let h be the fixed bandwidth used.

2. Then, set the bandwidth parameter $h(x_0)$ at location x_0 to $h(x_0) = h/\sqrt{\hat{K}(x_0)}$.

Intuitively, h controls the overall smoothing of the predictor space based on the sample points, and $\sqrt{\hat{K}(x_0)}$ stretches or shrinks the bandwidth at the particular location. The choice of the adaptive bandwidth parameter is motivated by the discussions of variable bandwidth kernel density estimators in Terrell & Scott (1992), Van Kerm (2003), and Abramson (1982), among others. In the case of d independent predictors, $h(\mathbf{x_0})$ at $\mathbf{x}_0$ is chosen as follows:

1. Compute the kernel density estimate $\hat{K}_i$, for $i \in 1, \cdots, d$, for the d predictors, with associated bandwidths $h_1, h_2, \cdots, h_d$, respectively. Then, h is chosen as the harmonic mean of h_i.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

2. Once h is obtained, the bandwidth parameter $h(\mathbf{x}_0)$ at $\mathbf{x_0}$ is given by

$$h(\mathbf{x}_0) = h/\left(\prod_{i=1}^{d}\sqrt{\hat{K}_i(x_{0i})}\right), \tag{5.1}$$

where x_{0i} is the ith coordinate of $\mathbf{x}_0$.

This choice of using the harmonic mean is based on the dependence of the minimax rates of convergence of the estimators to the harmonic mean of the smoothness of the density along the different dimensions, as discussed in Lepski (2015). We defer the analysis of the asymptotic properties of the proposed conditional density estimator to the Supplementary Material, Section S2.

5.1 Simulation Studies

Here, we present two examples to illustrate the proposed method and to compare its performance with that of the standard R package `NP` (using the kd-tree package implementation to reduce the computation time). In these experiments we have used a Gaussian family for f_p, the initial parametric conditional density estimate. To estimate the mean function, we have used a local linear regression function with Gaussian kernel weights, and with the bandwidth obtained using `kernel(bcv)`, available in the `R` package `kedd`. Bandwidths from other estimators, such as unbiased cross-validation, and even the naive *ksdensity* function

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

in MATLAB, produce qualitatively identical results. We use six basis elements for the tangent space representation throughout. Note that using Algorithm 1 to decide the number of basis elements would naturally increase the computation cost, depending on how many models are considered.

For comparison, we used 100 samples, each of size $n = 100$ and $n = 1000$, to obtain a mean integrated squared error loss function estimate, mean absolute error estimate, and mean $\mathbb{L}^\infty$ loss function estimate from the densities evaluated over a grid of 100 points at 10 equidistant locations over the support of each of the predictors. As a first example, we consider a situation where the true conditional density is a Laplace distribution; that is, $f(y_i|X = x_i) = \text{DExp}(y_i; \text{mean}=(2x_i - 1), \text{var}=1)$ and $X_i \sim \mathcal{N}(0, 1)$. As the second example, we consider a bivariate predictor scenario where $f(y_i|X = (x_{1i}, x_{2i})) = (1 - e^{-x_{2i}})\mathcal{N}(y_i; (x_{1i} + 2), (0.5)^2) + (e^{-x_{2i}})\text{DExp}(y_i; (x_{1i} - 1), 1)$, the predictors $X_1 \sim 0.95\mathcal{N}(0, (0.4)^2) + 0.05\mathcal{N}(0, (1.4)^2)$, and $X_2 \sim \mathbb{U}(0, 1)$.

The results are summarized in Table 3. From the results, it is clear that when the sample size is low, the performance of the warped estimate is better and more stable. When the sample size is high, the two methods perform similarly. For the second example, the NP package has better overall loss, although the warped estimation method still provides more stable performance. However, the computation cost of the NP package is very high, even with the kd-tree implementation,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

Table 3: *A comparison of the performance of the* `NP` *package and the warped estimate for the simulated examples.*

Method:			NP package			Warped Estimate		
Example	n	Norm	Mean	std.dev	Time	Mean	std.dev	Time
Example 1	100	$\mathbb{L}^1$	4.11	0.51		3.28	0.44	
		ISE	0.59	0.12	1 sec	0.41	0.11	1 sec
		$\mathbb{L}^\infty$	0.40	0.07		0.88	0.34	
	1000	$\mathbb{L}^1$	2.50	0.24		2.46	0.11	
		ISE	0.26	0.04	51 sec	0.25	0.03	3 sec
		$\mathbb{L}^\infty$	0.39	0.06		0.36	0.04	
Example 2	100	$\mathbb{L}^1$	60.49	6.67		58.55	5.28	
		ISE	11.43	4.01	2 sec	10.38	1.82	2 sec
		$\mathbb{L}^\infty$	2.47	0.43		2.41	0.35	
	1000	$\mathbb{L}^1$	42.10	4.32		53.53	1.86	
		ISE	5.88	1.41	198 sec	8.96	0.57	7 sec
		$\mathbb{L}^\infty$	2.38	0.29		2.24	0.25	

whereas the warped estimation is computationally very efficient.

5.2 Application to Epidemiology

Longnecker et al. (2001) studied the association between DDT metabolite DDE exposure and pre-term birth in a study based on the US Collaborative Perinatal Project (CPP). DDT is very effective against mosquitoes carrying malaria and, hence, is frequently used in malaria-endemic areas, in spite of evidence that suggests there are associated health risks. Both Longnecker et al. (2001) and Dunson & Park (2008) concluded that higher levels of DDE exposure are associated with higher risks of pre-term birth. The response variable in question is

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

the gestational age at delivery (GAD), and deliveries occurring prior to 37 weeks of gestation are considered as pre-term. Longnecker et al. (2001) also recorded the serum triglycerine level, among several other factors, and included it in their model, because the serum DDE level can be affected by the concentration of serum lipids.

We study the GAD data set to investigate the effects of varying levels of DDE on the distribution of the GAD, focusing on the left tail of the distribution to assess the effect on pre-term births. In our study, following Dunson & Park (2008), we include only the 2313 subjects for whom the gestation age at delivery is less than 45 weeks, attributing higher values to measurement errors. We study the conditional density of the GAD, given different doses of DDE in the serum. We also study the effects of different levels of triglyceride on the GAD. However, because DDE is a possible confounding factor, we conduct a bivariate analysis, including both DDE dose and triglyceride level as covariates, and study the effect on the GAD at varying levels of one covariate, keeping the other fixed. We also investigate whether different levels of one covariate affect the distribution of the other.

Based on our findings, the very erratic behavior at locations where the DDE dose or triglyceride levels lie in the 99th percentile is seen with some skepticism, owing to the sparsity of the data in that region. We notice an increasingly promi-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

nent peak near the left tail of the GAD distribution with increasing dose of DDE, which agrees with the results of Longnecker et al. (2001) and Dunson & Park (2008), shown in the left panel of Figure 3. The right panel of Figure 3 suggests a tendency of a higher risk of pre-term birth at higher doses of triglycerides as well, although the difference is less pronounced.

To investigate whether the results corresponding to triglycerides are confounded by the DDE doses, we first study the effects of triglyceride levels on the DDE distribution and vice versa. Figure 4 shows that the distributions of the covariates are almost identical for varying levels of the other. The only exception is at the 99th percentile of triglyceride, for which the distribution of the DDE doses seems to be shifted to the right. For fixed levels of triglyceride, increasing the DDE doses shows an increasing left peak, except where both the DDE and the triglyceride levels are very high, as shown in Figure 5. For fixed doses of DDE, the distribution of the GAD at different levels of triglyceride do not follow any increasing trend, and are almost indistinguishable from each other for different doses of DDE, as seen in Figure 6. This suggests that the increased risk of pre-term birth can be attributed primarily to DDE doses, and there is no significant effect of different triglyceride levels on the gestation age. The apparent increasing risk of pre-term birth for increasing level of triglycerides seen in the right panel of Figure 3 is caused mainly by DDE doses acting as a confounding

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

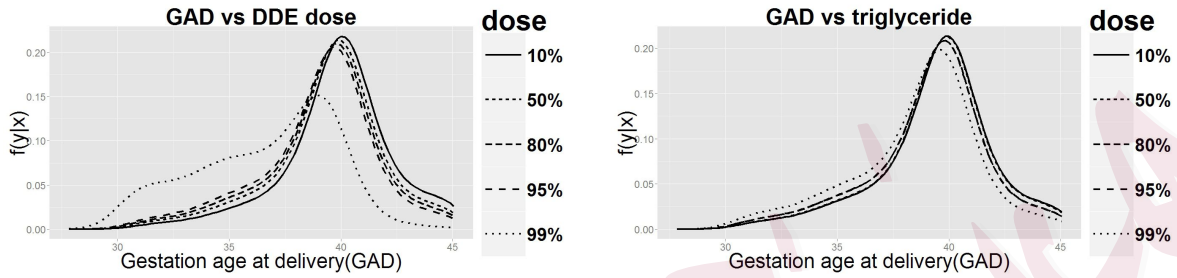

Figure 3: *Distribution of GAD for varying levels of DDE and triglyceride*

factor.

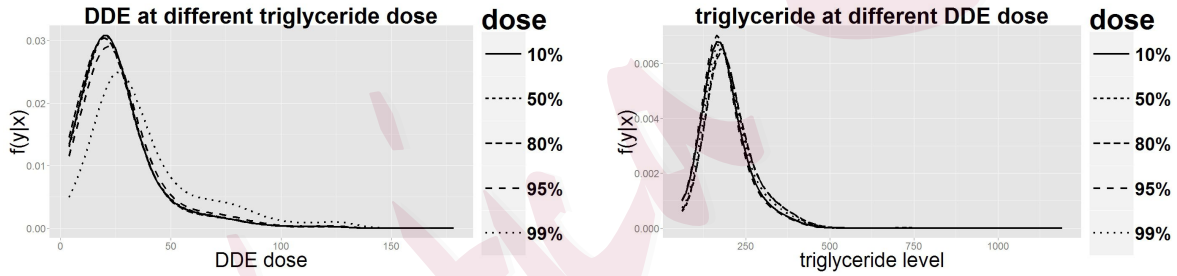

Figure 4: *Distribution of DDE and triglyceride at different levels of the other*

6. Discussion

Density estimation is a rich field of research in Statistics and machine learning. This study introduces a novel framework using geometric tools and the notion of a transitive group action, providing a new option for density estimation. Specifically, exploiting the geometry of the group of diffeomorphisms, we can shift the problem of finding an underlying density to one of finding an appropriate diffeo-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

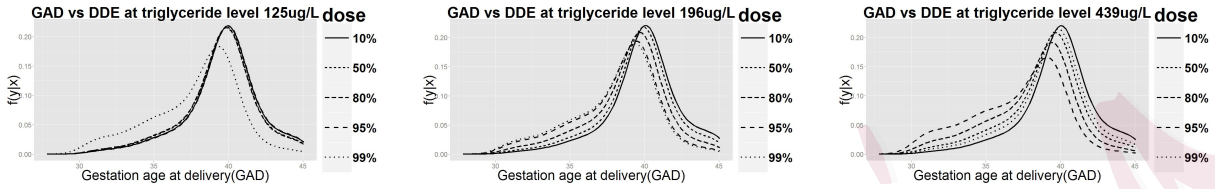

Figure 5: *Distribution of gestation at varying levels of DDE for fixed values of triglyceride*

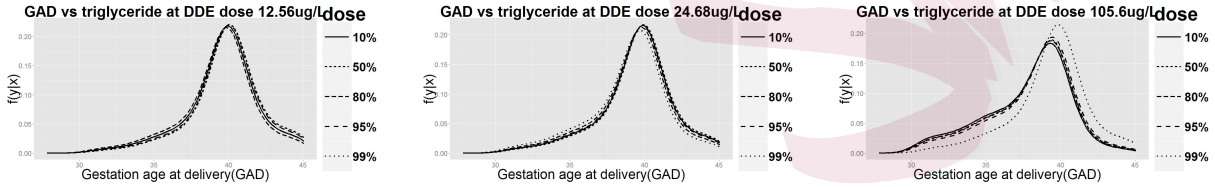

Figure 6: *Distribution of gestation at varying levels of triglyceride for fixed values of DDE*

morphism, given an initial shape, based on available data. In recent years, most data sets on a variable of interest have associated covariates that make a conditional density estimation useful and practically relevant. An advantage of the proposed framework is the easy extendibility of its geometric tools to the conditional density estimation problem, via a weighted maximum likelihood objective function.

Given the focus of our research, we touch only lightly upon, or do not explore many associated problems of density estimation, such as, the choice of the number of basis elements for the tangent space representation, choice of the basis set itself, or choice of a penalty for a penalized estimation and boundary

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

estimation. Here we use the AIC as the penalty to select the number of basis elements because we noticed that the BIC tends to choose an insufficient number of parameters. In addition, experiments using a Meyer basis set and a cosine basis set for the tangent space representation of the diffeomorphisms yielded similar results to those of the Fourier basis. Keeping in mind that the basis set representation is for used to approximate the warping functions and not the density functions directly, we can choose different basis sets for a comparative study of performances. Here, we follow Turnbull & Ghosh (2014) when choosing the boundaries.

For the conditional density estimation, the weights defined as a Gaussian kernel can also be defined using any other kernel. The choice of a Gaussian kernel (and the $\mathbb{L}^2$loss function) simply serves as an example. A possible extension is to estend the framework to include situations in which multiple or very high numbers of covariates are present. Currently, the bandwidth parameter is chosen adaptively based on a kernel density estimate at the location of the (scalar) covariate. It can be extended directly to d covariate scenario using a d variate kernel density estimate at the location of the predictors. However, such an estimate suffers from the curse of dimensionality. In applications where only a few of the covariates are relevant to the response variable, Wasserman & Lafferty (2006) developed a technique to identify the relevant variables, and to obtain the cor-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

responding bandwidth parameters. Using these bandwidth parameters, we can redefine the weights and perform a weighted likelihood maximization to produce a conditional density estimate. Furthermore, note that the proposed bandwidth for the weights in this study is same for all the covariates. However, many different isotropic and anisotropic data-driven bandwidth choices are possible, for example, using cross-validation, or the method proposed in Wasserman & Lafferty (2006). Note that even with an isotropic choice of bandwidth, the proposed technique performs similarly to a standard KDE, especially at smaller sample sizes.

Supplementary Material

The online Supplementary Material, Section S1, contains the proofs of the results in Section 3 of the manuscript. Section S2 discusses the asymptotic convergence rate of the conditional density estimator. Section S3 presents a practical implementation of the framework and several numerical techniques. Graphical representations of the univariate density estimations on the simulated data sets (see Section 4) are presented in Section S4 of the Supplementary Material. Section S5 presents simulation studies that investigate the properties of the density estimator. Section 6 discusses several properties of the estimator.

Acknowledgments

This research was supported in part by NSF grants to AS (NSF DMS

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

CDS&E 1621787 and NSF CCF 1617397). The authors would like to thank the anonymous referees and Associate Editor for their helpful comments.

References

Abramson, I. S. (1982), 'On bandwidth variation in kernel estimates-a square root law', *The annals of Statistics* pp. 1217–1223.

Bashtannyk, D. M. & Hyndman, R. J. (2001), 'Bandwidth selection for kernel conditional density estimation', *Computational Statistics & Data Analysis* **36**(3), 279–298.

Bhattacharya, A., Pati, D. & Dunson, D. (2010), 'Latent factor density regression models', *Biometrika* **97**(1), 1–7.

Birgé, L. & Massart, P. (1998), 'Minimum contrast estimators on sieves: exponential bounds and rates of convergence', *Bernoulli* **4**(3), 329–375.

Chung, Y. & Dunson, D. B. (2009), 'Nonparametric bayes conditional distribution modeling with variable selection', *Journal of the American Statistical Association* **104**(488), 1646–1660.

URL: *http://pubs.amstat.org/doi/abs/10.1198/jasa.2009.tm08302*

De Iorio, M., Muller, P., Rosner, G. L. & MacEachern, S. N. (2004), 'An anova model for dependent random measures', *Journal of the American Statistical Association* **99**(465), 205–215.

URL: *http://www.jstor.org/stable/27590366*

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1996), 'Density estimation by wavelet thresholding', *The Annals of Statistics* pp. 508–539.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

Doosti, H. & Hall, P. (2016), 'Making a non-parametric density estimator more attractive, and more accurate, by data perturbation', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**(2), 445–462.

Dunson, D. B. & Park, J.-H. (2008), 'Kernel stick-breaking processes', *Biometrika* **95**(2), 307–323.

Dunson, D. B., Pillai, N. & Park, J.-H. (2007), 'Bayesian density regression', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **69**(2), pp. 163–183.

URL: *http://www.jstor.org/stable/4623261*

Efromovich, S. (2010), 'Orthogonal series density estimation', *Wiley Interdisciplinary Reviews: Computational Statistics* **2**(4), 467–476.

Escobar, M. D. & West, M. (1995), 'Bayesian density estimation and inference using mixtures', *Journal of the american statistical association* **90**(430), 577–588.

Griffin, J. E. & Steel, M. F. J. (2006), 'Order-based dependent dirichlet processes', *Journal of the American Statistical Association* **101**(473), 179–194.

URL: *http://pubs.amstat.org/doi/abs/10.1198/016214505000000727*

Hall, P., Sheather, S. J., Jones, M. & Marron, J. S. (1991), 'On optimal data-based bandwidth selection in kernel density estimation', *Biometrika* **78**(2), 263–269.

Hansen, B. E. (2004), 'Nonparametric conditional density estimation', *Unpublished manuscript* .

Hjort, N. L. & Glad, I. K. (1995), 'Nonparametric density estimation with a parametric start', *The Annals of Statistics* pp. 882–904.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

Jain, S. & Neal, R. M. (2012), 'A split-merge markov chain monte carlo procedure for the dirichlet process mixture model', *Journal of Computational and Graphical Statistics* .

Kalli, M., Griffin, J. E. & Walker, S. G. (2011), 'Slice sampling mixture models', *Statistics and computing* **21**(1), 93–105.

Kundu, S. & Dunson, D. B. (2014), 'Latent factor models for density estimation', *Biometrika* **101**(3), 641–654.

Lang, S. (2012), *Fundamentals of differential geometry*, Vol. 191, Springer Science & Business Media.

Lenk, P. J. (1988), 'The logistic normal distribution for bayesian, nonparametric, predictive densities', *Journal of the American Statistical Association* **83**(402), 509–516.

Lenk, P. J. (1991), 'Towards a practicable bayesian nonparametric density estimator', *Biometrika* **78**(3), 531–543.

Leonard, T. (1978), 'Density estimation, stochastic processes and prior information', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 113–146.

Lepski, O. (2015), 'Adaptive estimation over anisotropic functional classes via oracle approach', *The Annals of Statistics* **43**(3), 1178–1242.

Li, Q. & Racine, J. S. (2007), *Nonparametric econometrics: theory and practice*, Princeton University Press.

Longnecker, M. P., Klebanoff, M. A., Zhou, H. & Brock, J. W. (2001), 'Association between maternal serum concentration of the ddt metabolite dde and preterm and small-for-gestational-age babies at

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

birth', *The Lancet* **358**(9276), 110–114.

MacEachern, S. N. (1999), 'Dependent nonparametric processes', *ASA Proceedings of the Section on Bayesian Statistical Science* .

URL: *http://aima.eecs.berkeley.edu/ russell/classes/cs294/f05/papers/maceachern-1999.pdf*

MacEachern, S. N. & Müller, P. (1998), 'Estimating mixture of dirichlet process models', *Journal of Computational and Graphical Statistics* **7**(2), 223–238.

Marron, J. S. & Wand, M. P. (1992), 'Exact mean integrated squared error', *The Annals of Statistics* pp. 712–736.

Müller, P., Erkanli, A. & West, M. (1996), 'Bayesian curve fitting using multivariate normal mixtures', *Biometrika* **83**(1), 67–79.

Norets, A. & Pelenis, J. (2012), 'Bayesian modeling of joint and conditional distributions', *Journal of Econometrics* **168**, 332–346.

Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function', *The Annals of Mathematical Statistics* **27**(3), 832–837.

Saoudi, S., Ghorbel, F. & Hillion, A. (1997), 'Some statistical properties of the kernel-diffeomorphism estimator', *Applied stochastic models and data analysis* **13**(1), 39–58.

Saoudi, S., Hillion, A. & Ghorbel, F. (1994), 'Non–parametric probability density function estimation on a bounded support: Applications to shape classification and speech coding', *Applied Stochastic models and data analysis* **10**(3), 215–231.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

Sheather, S. J. & Jones, M. C. (1991), 'A reliable data-based bandwidth selection method for kernel density estimation', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 683–690.

Srivastava, A. & Klassen, E. P. (2016), *Functional and shape data analysis*, Springer.

Tabak, E. G. & Trigila, G. (2014), 'Data-driven optimal transport', *Commun. Pure. Appl. Math. doi* **10**, 1002.

Tabak, E. & Turner, C. V. (2013), 'A family of nonparametric density estimation algorithms', *Communications on Pure and Applied Mathematics* **66**(2), 145–164.

Terrell, G. R. & Scott, D. W. (1992), 'Variable kernel density estimation', *The Annals of Statistics* pp. 1236–1265.

Tokdar, S. T. (2007), 'Towards a faster implementation of density estimation with logistic gaussian process priors', *Journal of Computational and Graphical Statistics* **16**(3), 633–655.

Tokdar, S. T., Zhu, Y. M. & Ghosh, J. K. (2010), 'Bayesian density regression with logistic gaussian process and subspace projection', *Bayesian analysis* **5**(2), 319–344.

Turnbull, B. C. & Ghosh, S. K. (2014), 'Unimodal density estimation using bernstein polynomials', *Computational Statistics & Data Analysis* **72**, 13–29.

Van Kerm, P. (2003), 'Adaptive kernel density estimation', *Stata Journal* **3**(2), 148–156.

Vermehren, V. & de Oliveira, H. (2015), 'Close expressions for meyer wavelet and scale function', *arXiv preprint arXiv:1502.00161* .

Wasserman, L. & Lafferty, J. D. (2006), Rodeo: Sparse nonparametric regression in high dimensions, *in*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0231

'Advances in Neural Information Processing Systems', pp. 707–714.

Wong, W. H. & Shen, X. (1995), 'Probability inequalities for likelihood ratios and convergence rates of sieve mles', *The Annals of Statistics* pp. 339–362.

Florida State University

E-mail: (s.dasgupta@stat.fsu.edu)

Texas A&M University

E-mail: (debdeep@stat.tamu.edu)

Florida State University

E-mail: (anuj@stat.fsu.edu)