QUARTERLY OF APPLIED MATHEMATICS
VOLUME LXXVII, NUMBER 2
JUNE 2019, PAGES 399–422
https://doi.org/10.1090/qam/1529
Article electronically published on January 8, 2019

BAYESIAN SHAPE-CONSTRAINED DENSITY ESTIMATION

By

SUTANOY DASGUPTA (*Department of Statistics, Florida State University, Tallahassee, Florida 32036*),

DEBDEEP PATI (*Department of Statistics, Texas A&M University, College Station, Texas 77843*),

AND

ANUJ SRIVASTAVA (*Department of Statistics, Florida State University, Tallahassee, Florida 32036*)

This paper is dedicated to Professor Ulf Grenander

Abstract. The problem of estimating probability densities underlying given *i.i.d.* samples is a fundamental problem in statistics. Taking a Bayesian nonparametric approach, we put forth a geometric solution that uses different actions of the diffeomorphism (domain warping) group on the set of positive *pdf*s to explore this space more efficiently. This representation shifts the focus from *pdf*s to the diffeomorphism group and allows efficient solutions for density estimation under shape (or modality) constraints, i.e., estimation of a *pdf* given a fixed or a maximum number of modes. Focusing on univariate density estimation, we use the geometry of a (one-dimensional) diffeomorphism group to reach an (approximate) finite-dimensional Euclidean representation of warping functions, and impose a shrinkage prior on this space to form a posterior distribution. We sample this posterior using the Markov Chain Monte Carlo algorithm and form Bayesian estimates of the unknown *pdf*. This framework results in a novel *pdf* estimator, with and without shape constraints, and we demonstrate it in a number of simulated and real data experiments.

Received April 16, 2018, and, in revised form, October 8, 2018.
2010 *Mathematics Subject Classification.* Primary 65C60, 62G05; Secondary 57N25, 49Q10.
Key words and phrases. Shape analysis, density estimation, warping groups, deformable template, shape constraints.
The second author's research was supported by NSF DMS 1613156.
The third author's research was supported in part by the NSF grants to AS – NSF DMS CDS&E 1621787 and NSF CCF 1617397.
Email address: sdasgupta@stat.fsu.edu
Email address: debdeep@stat.tamu.edu
Email address: anuj@stat.fsu.edu

Contents

1. Introduction. Professpr Ulf Grenander was an exceptional researcher, scientist, and mathematician of his generation, a pioneer in many ways. His expertise spanned a wide swath of topic areas, with focus on both classical and modern computational statistics. Not surprisingly, he is credited with major innovations in statistics and engineering. He is greatly admired for his bold initiatives, sometimes without a full support from his contemporary researchers, along directions that proved fruitful many years later. These initiatives have had profound impacts on the community, and papers are still being written on those topics, even several decades after his original work. Indeed these ideas continue to motivate research and development for new and emerging types of datasets. One such area is **shape-constrained** density estimation, where Grenander's estimator [14] stands as a foundational result on which a whole research community has been developed. While his work, and the work that followed, focused on the estimation of densities with very specific shapes—*unimodal* and *monotone* densities—it seems useful to generalize this idea to a more general notion of shape. Another of Grenander's pioneering contributions was his formalization of the **deformable template** theory [1, 15], in which one reaches a large span of objects by deforming a prototype or a template. This representation is based on the action of the deformation groups on the space of complex objects, and the main advantage here is to transfer the process of inference from a complex space of objects to a much simpler group of deformations. One can use known geometries of these deformation groups to impose distributions and derive inferences on these structured spaces.

In this paper we combine these two of Grenander's seminal contributions—*shape-constrained* density estimation and *deformable template* theory—to develop geometric, Bayesian approaches to nonparametric density estimation. The problem of estimating probability density functions (*pdfs*) from their samples is a classical problem in statistics and has been studied for several decades in a variety of contexts. Consequently, a wide variety of solutions have been developed, each with their own strengths and limitations. The broad taxonomy of these solutions are usually parametric or nonparametric, frequentist or Bayesian, constrained or unconstrained, and so on. While early research favored parametric solutions, in view of their simplicity and analyzability, the focus has recently shifted to nonparametric solutions, especially under a Bayesian paradigm. Even though these solutions represent tremendous progress in the field, there is still plenty of scope for developments especially for computationally efficient solutions involving large datasets.

Recent years have seen great advances in the area of functional data analysis, especially using geometric approaches. The key idea is to consider functions as elements of a certain functional space, rather than as a set of scalar values over time intervals. This enables one to exploit the geometry of functional spaces and to develop efficient algorithms for inference under such geometric representations. Viewing *pdf*s as functions on a fixed domain, such as an interval or a unit square or a unit sphere, one can develop a new perspective on density estimation and statistical inferences. The estimation framework can either be frequentist or Bayesian, but it is framed atop a mathematical representation of *pdf*s that is derived from a geometrical perspective. In this paper we develop a Bayesian point of view and demonstrate the advantages of a geometric approach over conventional solutions.

Consider the following setup: Given a set of samples $\{x_i \in D, i = 1, 2, \ldots, n\}$ from a density function f_0 on a compact domain D, our goal is to estimate f_0. In this context, we consider two problems:

(1) **Problem 1: Unconstrained Density Estimation**. Suppose there is an efficient technique, e.g., a parametric solution, to provide a rough estimate f_p of f_0. How can one refine this initial estimate to reach an optimal solution while retaining efficiency? The idea is to bridge the gap between these two densities, f_p and f_0, using deformable template theory, and to implement it using a fast algorithm.

(2) **Problem 2: Shape-Constrained Density Estimation**. In case we know that f_0 is of certain shape—for example, we know either the precise number or maximum number of modes of f_0—then, how can we incorporate that shape information in an estimation of f_0? The goal here is to identify the set of valid densities and to define an optimal solution in this set according to some chosen criterion. The construction of a computationally efficient estimator requires the ability to explore the constraint set without venturing into the larger set of all *pdf*s.

A common approach for solving Problem 1 is a two-step estimation procedure discussed in [25–27, 39, 40], and some others. Here one *improves* upon an initial rough guess f_p by forming a function $w > 0$, that depends on the initial estimate f_p, and obtaining

a final estimate $wf_p / \int_y wf_p \, dy$. Thus, the second step involves the estimation of an optimal w in order to reach the final estimate. In a Bayesian context, the function w is often previously assigned a Gaussian process [25, 39, 40]. While this approach is quite comprehensive, the calculation of the normalization constant at every iteration makes computations very cumbersome.

Solutions to Problem 2 are not readily available in the literature in general. For certain simple constraints, such as monotonicity, unimodality, and log concavity, there have been extensive studies in the past, but for more general constraints, such as M-modality constraints, where the density f_0 is fixed to have a certain $M > 0$ modes, the literature is essentially wide open. There are plenty of instances in social, economical, and natural sciences where the distributions naturally attain multimodal structure. Examples include intensities of growth spurts, age stamps of disease occurrences in humans, colors of galaxies, electricity consumption profiles of households, and so on. Similarly, there are also plenty of instances of general functions (see [23] for example) that exhibit shapes other than monotonicity and unimodality. Unconstrained estimators that usually minimize the mean squared error tend to overestimate the number of modes for small or moderate sample sizes. Ensuring that an estimate lies in the correct shape class for any sample size lends interpretability to the observed features of the estimate. Moreover, obtaining an "optimal" estimate in the correct shape class makes the estimate more reliable than an ad hoc choice within the shape class.

We investigate both of these problems using a geometric approach that relies on the action of a certain deformation group on the density spaces of interests. In view of this group action, the problem of inference transfers from density spaces to the group. Utilizing the differential geometry of a deformation group, we perform optimizations over all deformations to perform density estimation and develop techniques for Bayesian nonparametric inferences. Although we will illustrate the solutions using univariate density estimation, for simplicity of computation and explanation, these methods also naturally apply to any higher-dimensional domain.

1.1. *Past research.* We summarize past research from different domains that relate to the two problems posed in the previous section.

- **Unconstrained Density Estimation**: Nonparametric solutions, especially kernel based estimators, are currently the norm in the literature for unconstrained density estimation. Please refer to [17, 28, 34, 35] for a narrative on this framework. Related to these approaches are "tilting" or "data sharpening" techniques for unconditional density estimation; see for example [11, 19] and the references therein. Bayesian approaches provide good solutions to the problem albeit at the cost of high computational complexity. Faster computational solutions, beyond the classical Metropolis Hastings implementations, have become popular in the Bayesian nonparametric community. Over the recent years, Bayesian methods for estimating *pdfs* based on mixture models and latent variables have received a lot of attention, primarily due to their excellent practical performances and an increasingly rich set of algorithmic tools for sampling from posterior distributions using Markov Chain Monte Carlo (MCMC) methods. References include [4, 12, 21, 22, 24, 29, 32] and many others.

- **Shape-Constrained Density Estimation**: In the context of shape-constrained density estimation, there is an extensive literature on analysis and modifications for the Grenander estimator for monotonic and unimodal density estimators. Rao [33] and others [16, 43] studied asymptotic properties of Grenander's estimator and established its consistency. Others [6, 8, 30] extended this framework to include nonsmooth unimodal densities. More recently, a number of papers including [5] have broadened the available tools for unimodal density estimation. Izenman [20] provides a review of nonparametric approaches, including those estimating unimodal densities. A more recent paper by Turnbull et al. [41] uses the Bernstein polynomial basis for estimating unimodal densities from the data.

  The literature on shape-constrained estimation dates back at least to the 1950s ([14, 18]). For monotone or concave/convex constraints, the classical constrained least squares estimator discussed in [18] minimizes a least squares criterion subject to the said constraints. Various theoretical properties of the least squares estimator, including consistency, rates of convergence, and asymptotic distribution, have been derived, as in [2, 7]. More recent theoretical contributions include extensions of these results to the multivariate setting ([13]) and deriving rates and sharp oracle inequalities under minimum smoothness assumptions, such as Lipschitz continuity ([3]). While these are impressive theoretical results, it is often the case in scientific and engineering applications that one needs to impose constraints on the shape of the density and to characterize uncertainty in estimating the density. While bootstrapping is a popular approach to characterize uncertainty, the theory for such estimators is not as well-developed, especially in terms of uncertainty characterization. In this article, we develop a fully probabilistic way of learning the density subject to a constraint on the number of modes.
- **Shape Analysis of Elastic Euclidean Curves**: A topic related to Problem 2 is the shape analysis of Euclidean curves, of the type $\beta : [0, 1] \to \mathbb{R}^n$, where one studies shapes formed by such curves. Shape is a property that is invariant to certain transformations, such as rigid motions, global scaling, and reparameterizations. Over the last decade there have been several important developments in this area, including: (1) introductions of elastic Riemannian metrics that invariant to the actions of these shape-preserving transformations, and (2) certain square-root transformations that transform these complex metrics into more practical $\mathbb{L}^2$ metrics. For some general discussion on this topic, please refer to [31, 37, 38, 45, 46]. If we set $n = 1$, then shape analysis becomes related to the current problem of shape-constrained density estimation. In fact, the tools derived in the current paper can be extended to the estimation of Euclidean curves under given shape constraints. However, we have not studied this direction in this paper and have left it out for future research.

1.2. *Proposed geometric framework.* Dasgupta et al. [10] and [9] introduced a geometric approach for solving Problems 1 and 2, and estimated densities by *deforming* an initial *pdf* into an optimal solution. The deformation, in turn, is based on the action of a group of deformations of the domain D, with the action chosen according to the need

of the problem. (In the univariate case these domain deformations are often known as *time warpings*.) An area-preserving action helps search over all *pdf*s, in an unconstrained way, by varying the deformations and steering the solution towards the optimal *pdf*. A shape-preserving or a mode-preserving group action helps search for optimal solutions under the constraints of known shapes. This framework shifts the burden of estimation or optimization onto the group of deformations. Using the geometries of deformation groups, one can efficiently reach optimal estimates. However, [10] and [9] take a frequentist approach and rely on inbuilt `MATLAB` functions for log-likelihood optimization. As a result there is little to no control on the optimization procedure itself. A Bayesian approach on the other hand provides a lot more control over the optimization and flexibility in the search through the choice of the proposal density and the choice of the prior structure. Furthermore, the Bayesian approach naturally allows one to gauge the uncertainty in the estimate through the posterior credible intervals. Hence in this paper we introduce a Bayesian and geometric approach to density estimation, and investigate its strength and weaknesses. We define prior distributions on the set of deformations and seek posterior samples for this set to form density estimates.

The novel contributions of this paper are:

(1) It provides a unified discussion on unconstrained and constrained density estimation using different actions of a deformation group to explore the space of all *pdf*s.
(2) It introduces a novel Bayesian framework to represent and search over the deformation group utilizing the underlying geometry of the group.
(3) It provides an efficient Bayesian density estimation under proper multimodal (shape) constraints.
(4) It demonstrates these ideas using simple examples from univariate density estimation and compares the performance of different prior structures.

The rest of this paper is organized as follows. In Section 2, we introduce different actions of the diffeomorphism groups on the space of positive density functions on a compact domain D. In Section 3, we prove some background material on geometries of relevant spaces. In Sections 4 and 5, we present framework for unconstrained and constrained density estimation, respectively, and in Section 6 we present Bayesian solutions to these problems. We present a real data analysis in Section 7, involving the electricity consumption pattern in a random household in Tallahassee. We finish the paper with a simple discussion in Section 7.

2. Deformation groups, pdfs, and shapes. The approach pursued in this paper is to focus on the geometry of the space of *pdf*s and how to explore it efficiently, with or without any constraints. These explorations, in turn, lead to desired *pdf* estimators. The main tool used in traversing the space of *pdf*s is using a suitable action of the diffeomorphism group on that space, termed as *deformations*. These actions primarily deform current *pdf*s by warping the domain D on which the densities are defined, and adapt the heights accordingly, to help us traverse the density function space. There are different actions enabling explorations suitable for different situations.

Let Γ denote the set of all orientation-preserving diffeomorphisms from a domain D to itself. Γ is an infinite-dimensional Lie group with the group operation being composition $\circ$ and the identity element $\gamma_{id}(s) = s$. In case D is one-dimensional, then Γ is also referred to as the *time-warping group* or simply the *warping* group. Let $\mathcal{F}$ be the set of all smooth functions on D; Γ acts on $\mathcal{F}$ in a number of ways, and at least two of them will be useful in density estimation.

(1) **Area-Preserving**: This action preserves the area below the curve in a function and is a mapping from $\mathcal{F} \times \Gamma \to \mathcal{F}$ given by $(f, \gamma) = (f \circ \gamma)J_\gamma$. Here J_γ denotes the determinant of the Jacobian of the diffeomorphism $\gamma : D \to D$. It is easy to verify that $\int_D f(s)\ ds = \int_D (f, \gamma)\ ds$. This mapping is akin to the change of variables formula for random variables. This action is especially suitable for probability densities since a *pdf* remains a *pdf* under this mapping. However, this mapping can change the *shape* of a function by changing the relative heights of its peaks and valleys, introducing new modes or removing modes, and so on. Also, the $\mathbb{L}^2$ norm of a function is not preserved, i.e., $\|f\| \neq \|(f \circ \gamma)J_\gamma\|$ for a general f and γ. As a special case, for $D = \mathbb{R}$ or $[0, 1]$, we have $(f, \gamma) = (f \circ \gamma)\dot{\gamma}$.

(2) **Shape-Preserving**: This action is based on a simple deformation of the domain D by the group Γ. It is given by the mapping from $\mathcal{F} \times \Gamma \to \mathcal{F}$ defined as $(f, \gamma) = f \circ \gamma$, where $\circ$ denotes a composition. An important property of this action is that all the heights in f are preserved in the mapping from f to $(f \circ \gamma)$; they simply get shifted horizontally according to γ. Thus, in a sense, it preserves the *shape* of a function. Shape in this context signifies the number of modes and antimodes of a density function. However, the area below a curve can change under this group action since $\int_D f(s)\ ds \neq \int_D f(\gamma(s))\ ds$ in general. Also, the $\mathbb{L}^2$ norm of a function is not preserved, i.e., $\|f\| \neq \|(f \circ \gamma)\|$ for a $f \in \mathcal{F}$ and $\gamma \in \Gamma$ in general. One can easily modify this action (by adding a normalization) to make this action area-preserving: $(f, \gamma) = \frac{(f \circ \gamma)}{\int_D (f \circ \gamma) ds}$. This last action is thus both shape- (or mode-) and area-preserving.

(3) **Norm-Preserving**: This action preserves the $\mathbb{L}^2$ norm of a function. It is defined as a mapping from $\mathcal{F} \times \Gamma \to \mathcal{F}$ given by $(f, \gamma) = (f \circ \gamma)\sqrt{J_\gamma}$. It can be verified that $\|f\| = \|(f \circ \gamma)\sqrt{J_\gamma}\|$ for all $f \in \mathbb{L}^2$ and $\gamma \in \Gamma$. However, this action does not preserve the shape of f, as the peaks and valleys can added or removed under this mapping. It also does not preserve the area below the curve, and therefore the resulting function may not be a *pdf*. As a special case, for $D = \mathbb{R}$ or $[0, 1]$, we have $(f, \gamma) = (f \circ \gamma)\sqrt{\dot{\gamma}}$.

Figure 1 shows an example of each of the three actions of a univariate Γ on a *pdf* on the domain $D = [0, 1]$. The top row shows the original *pdf* f and the three time warping functions $\{\gamma_i, i = 1, 2, 3\}$. The bottom row shows the three actions listed above. It can be seen that while the original f is bimodal, the area- and the norm-preserving actions can result in trimodal structures. On the other hand, all the deformed functions are still bimodal under mode-preserving transformations.

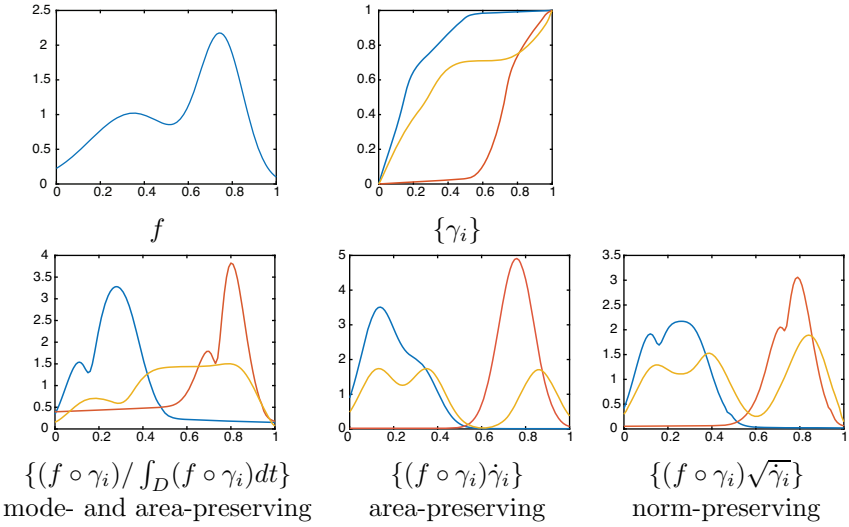

FIG. 1. Top row: The left panel shows a density f that is deformed using a number of warping functions (middle panel). Bottom row: The different actions of the warping functions resulting in proper *pdf*s shown in the left and middle panel because the transformation is area-preserving. The right panel shows functions which are not proper densities.

Figure 2 shows an example of bivariate density on the domain $D = [0,1]^2$. The top-left panel in this figure shows a bimodal pdf f that is acted upon by a diffeomorphism γ in two different ways. We display this γ using its displacement field $\gamma - \gamma_{id}$ (top-middle panel) and its determinant of the Jacobian J_γ (top-right panel). The bottom row shows the two group actions: (1) mode-preserving action $\{(f\circ\gamma)/\int_D(f\circ\gamma)ds\}$ and (2) area-preserving action $\{(f\circ\gamma)J_\gamma\}$. It can be seen that the first action only moves the modes horizontally and scales them globally, but the second action also changes the relative heights of the two peaks.

The first two actions are useful for exploring the space of density functions and reaching efficient estimators, depending upon the context. In case one wants to estimate a density without any constraint, then the area-preserving action is a suitable choice. One can start with an arbitrary initial estimate of a *pdf* and search over all *pdf*s by applying different elements of Γ. Instead, if one is interested in estimating a *pdf* with a certain shape, then it is better to use the shape-preserving transformation. In this case, one can start with an arbitrary element of the correct shape class and then search over that shape class by optimizing over elements of Γ applied to that original element.

3. Background material of geometry.

3.1. *Geometry of a pdf space.* Before we present the main idea, we briefly describe some basic geometry of the set of *pdf*s. Let $\mathcal{P}$ denote the set of all strictly positive *pdf*s on a compact domain D. $\mathcal{P}$ is a Banach manifold with the ambient space for coordinate charts coming from the space of $\mathbb{L}^1$ functions on D. As stated earlier, our

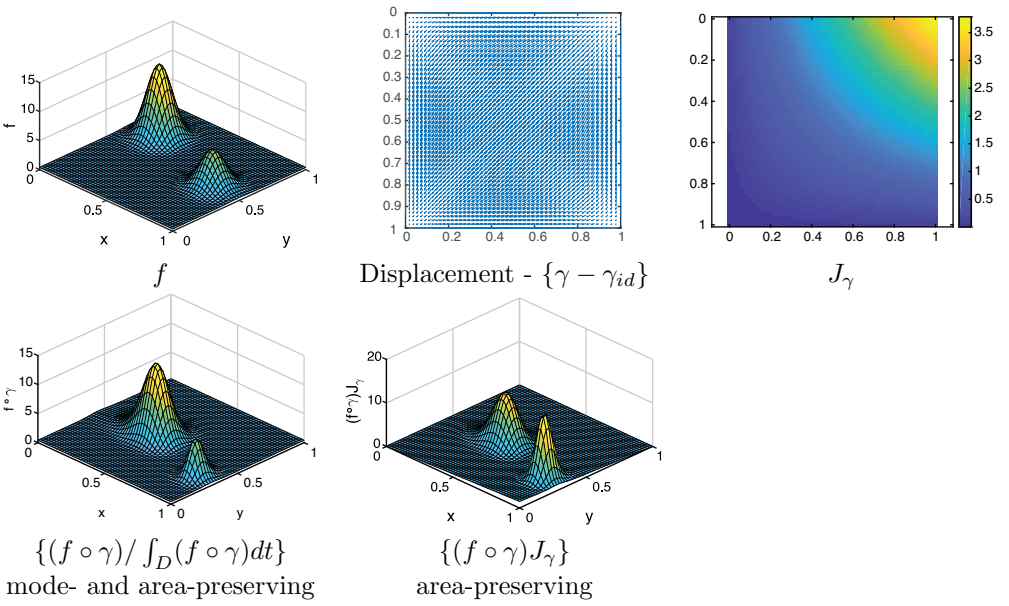

FIG. 2. Top row: The left panel shows a density f that is deformed using a diffeomorphism γ. We display the corresponding displacements $\gamma - \gamma_{id}$ (middle), and determinant of the Jacobian J_γ (right). Bottom row: The different actions of the diffeomorphism resulting in proper *pdf*s shown in the left and middle panel because the transformation is area-preserving.

approach is to use geometry of $\mathcal{P}$ for purposes of estimation and inference of *pdf*s from sampled data. The set $\mathcal{P}$ in itself has the following interesting geometry. For any element $g \in \mathcal{P}$, its positive square root $q(s) = \sqrt{g(s)}$ is an element of the unit Hilbert sphere $\mathbb{S}_\infty = \{g : D \to \mathbb{R} | \int_D g(t)^2 dt = 1\}$, because $\int_D q^2(s)\, ds = \int_D g(s)\, ds = 1$. The natural distance for measuring differences between elements of $\mathbb{S}_\infty$ is the *arc length* on $\mathbb{S}_\infty$; this originates from the standard $\mathbb{L}^2$ Riemannian metric (inner product) on tangent spaces of $\mathbb{S}_\infty$. That is, for any $g_1, g_2 \in \mathbb{S}_\infty$, $d_g(g_1, g_2) = \cos^{-1}(\langle g_1, g_2 \rangle)$. This is called the *Fisher–Rao* distance between *pdf*s. While one can exploit this spherical geometry to derive better density estimators, even more flexibility and efficiency can be obtained using deformations as described next.

3.2. *Geometry of a diffeomorphism group.* Recall that Γ is a set of all orientation-preserving diffeomorphisms from a domain D to itself. The set Γ is a Lie group with the group operation given by composition: for any $\gamma_1, \gamma_2 \in \Gamma$, the group operation is $(\gamma_1 \circ \gamma_2)(s) = \gamma_1(\gamma_2(s))$. The identity element of Γ is the identity map: $\gamma_{id}(s) = s$ and for any $\gamma \in \Gamma$, its inverse γ^{-1} is well defined such that $\gamma \circ \gamma^{-1} = \gamma^{-1} \circ \gamma = \gamma_{id}$.

While Γ is a group, it is not a vector space; a linear combination of elements of Γ may not be in Γ. (However, Γ is closed under *convex* combinations, i.e., linear combinations where the coefficients add up to 1 and are all nonnegative.) The tangent structure of Γ at γ_{id}, $T_{\gamma_{id}}(\Gamma)$, is the set of smooth vector fields that are tangential to the boundaries at

the boundary points:

$$T_{\gamma_{id}}(\Gamma) = \{v : D \mapsto \mathbb{R}^{\dim(D)} | v(\delta D) = T(\delta D), \quad v \text{ is smooth}\} \ .$$

Here $T(\delta D)$ denotes the tangent bundle of the boundary δD. The tangent space $T_{\gamma_{id}}(\Gamma)$ is a vector space although not a Hilbert space.

In the deformable template representation pursued in this paper, the problem of inference lies on Γ. Taking a Bayesian approach, we impose prior distributions on this space and seek samples from the resulting posterior. This task is complicated due to the nonlinear and infinite-dimensional nature of Γ. Therefore, a further understanding of the geometry of Γ can help in this regard. One can make Bayesian explorations more efficient by using elements of this geometric structure.

3.3. *Univariate diffeomorphisms.* Consider the special case when $D = [0, 1]$ and Γ is the set of univariate diffeomorphisms of $[0, 1]$ to itself. The tangent space in this case is

$$T_{\gamma_{id}}(\Gamma) = \{v : [0, 1] \mapsto \mathbb{R} | v(0) = 0, v(1) = 0, \quad v \text{ is smooth}\} \ .$$

We are interested in solving optimization problems on Γ, and some finite-dimensional approximations of elements of $T_{\gamma_{id}}(\Gamma)$ are very useful. Towards that goal, we present an orthonormal basis of this space that can be conveniently truncated for finite-dimensional approximations. This basis representation requires a metric to specify orthogonality, and we will use the Fisher–Rao metric for this purpose. For any $v_1, v_2 \in T_\gamma(\Gamma)$, it takes the form

$$\langle\langle v_1, v_2 \rangle\rangle_\gamma = \int_0^1 \dot{v}_1(s)\dot{v}_2(s)\frac{1}{\dot{\gamma}(s)}ds \ . \tag{3.1}$$

For a $\gamma \in \Gamma$, the derivative $\dot{\gamma}$ is an element of $\mathcal{P}$ defined in the previous section. Therefore, we have a natural mapping $\gamma \to \dot{\gamma} \to \sqrt{\dot{\gamma}}$, termed *SRVF*, from Γ to $\mathcal{P}$ to $\mathcal{Q}$, where

$$\mathcal{Q} = \{q : D \to \mathbb{R}_+ | \int_D q(s)^2 \ ds = 1\} \ .$$

As described in [38], the Fisher–Rao metric under SRVF transforms to the $\mathbb{L}^2$ Riemannian metric on $\mathcal{Q}$, $\mathcal{Q}$ is termed a positive orthant of the Hilbert sphere $\mathbb{S}_\infty$, and the geodesics on $\mathcal{Q}$ are simply arcs on great circles. An orthogonal basis of $T_{\gamma_{id}}(\Gamma)$, under the Fisher–Rao metric, can be easily written using the geometry of $\mathbb{S}_\infty$. Since γ_{id} maps to a constant function $\mathbf{1}$, the tangent space $T_{\mathbf{1}}(\mathbb{S}_\infty) = \{w : [0, 1] \to \mathbb{R} | w \text{ is smooth}, \int_0^1 w(s)ds = 0\}$. One can impose a natural Hilbert structure on $T_{\mathbf{1}}(\mathbb{S}_\infty)$ using the standard inner product: $\langle w_1, w_2 \rangle = \int_0^1 w_1(s)w_2(s)ds$. Elements of this space can be mapped back to $\mathbb{S}_\infty$ via a *retraction* and subsequently to $T_{\gamma_{id}}(\Gamma)$ using the square-integral mapping mentioned above, illustrated in Figure 3. This *retraction* is carried out using the exponential map:

$$\exp_{\mathbf{1}}(w) : T_{\mathbf{1}}(\mathbb{S}_\infty) \to \mathbb{S}_\infty, \quad \exp_{\mathbf{1}}(w) = \cos(\|w\|)\mathbf{1} + \frac{\sin(\|w\|)}{\|w\|} \ . \tag{3.2}$$

Thus, any orthonormal basis of $T_{\mathbf{1}}(\mathbb{S}_\infty)$ under the $\mathbb{L}^2$ metric results in, through this mapping, an orthonormal basis for $T_{\gamma_{id}}(\Gamma)$ under the Fisher–Rao metric.

So far we have found a mapping between the elements of Γ and $T_{\mathbf{1}}(\mathbb{S}_\infty)$, and this helps deal with the nonlinearity of Γ, since $T_{\mathbf{1}}(\mathbb{S}_\infty)$ is a Hilbert space (with the standard $\mathbb{L}^2$

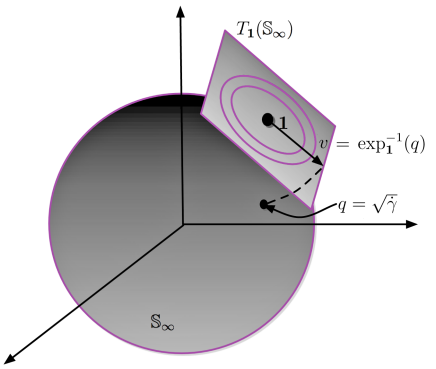

FIG. 3. Representing warping function γ as element of the tangent space $T_{\mathbf{1}}(\mathbb{S}_\infty^+)$.

inner product). To find an (approximate) finite-dimensional representation of elements of Γ, one can choose any orthonormal basis of $T_{\mathbf{1}}(\mathbb{S}_\infty)$ and truncate it using some criterion. Then, the coefficients of the basis elements (given a fixed basis set) behave as Euclidean parameters which uniquely describe a $\gamma \in T_{\mathbf{1}}(\mathbb{S}_\infty)$. Given a basis set $\mathcal{B} = \{b_j, j = 1, 2, \dots\}$ and a truncation to J basis elements, this mapping from the set of coefficients to a warping function can be summarized through a composite map $H : \mathbb{R}^J \to \Gamma$, given by

$$\{c_j\} \in \mathbb{R}^J \xrightarrow{\{b_j\}} w = \sum_{j=1}^{J} c_j b_j \in T_{\mathbf{1}}(\mathbb{S}_\infty) \xrightarrow{\exp_{\mathbf{1}}} q \in \mathbb{S}_\infty \to \gamma(t) = \int_0^t q(s)^2 ds \,. \quad (3.3)$$

4. Problem 1: Unconstrained density estimation. In this section we develop a two-step framework for estimating unconditional *pdf*, and start by introducing some notation. Let $\mathcal{F}$ be the set of all strictly positive, probability density functions on D. Let $f_0 \in \mathcal{F}$ denote the underlying true density, and let $x_i \sim f_0$, $i = 1, 2, \dots, n$ be independent samples from f_0. Furthermore, let $\mathcal{F}_p$ be a predetermined subset of $\mathcal{F}$ such that an optimal element (based on likelihood or any other desired criterion) $f_p \in \mathcal{F}_p$ is relatively easy to compute. For instance, any parametric family with a simple maximum-likelihood estimator is a good candidate for f_p. Similarly, kernel density estimates are also good since they are computationally efficient and robust in univariate setups.

Next, we define a warping-based transformation of elements of $\mathcal{F}_p$, using elements of Γ defined earlier. For any $f_p \in \mathcal{F}_p$ and $\gamma \in \Gamma$, define the mapping $(f_p, \gamma) = (f_p \circ \gamma) J_\gamma$, previously called area-preserving action. In the univariate setting, this mapping reduces to $(f_p, \gamma) = (f_p \circ \gamma)\dot{\gamma}$. The importance of this mapping comes from the following result.

PROPOSITION 4.1.

(1) The mapping $\mathcal{F} \times \Gamma \to \mathcal{F}$, specified above, forms an action of Γ on $\mathcal{F}$.

(2) In a univariate setting, this action is transitive. In other words, one can reach any element of $\mathcal{F}$ from any other element of $\mathcal{F}$ using an appropriate element of Γ.

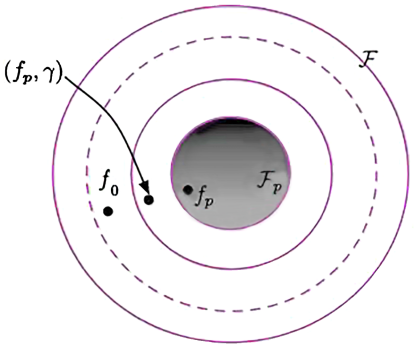

FIG. 4. The true pdf f_0 is estimated by transforming an initial estimate f_p by the warping function γ. The larger the set of allowed γs, the better the estimate.

Proof.

(1) We can verify the two properties in the definition of a group action: (i) For any $\gamma_1, \gamma_2 \in \Gamma$ and $f \in \mathcal{F}$, we have $((f, \gamma_1), \gamma_2) = (((f \circ \gamma_1)J_{\gamma_1}) \circ \gamma_2)J_{\gamma_2} = (f, \gamma_1 \circ \gamma_2)$. (ii) For any $f \in \mathcal{F}$, $(f, \gamma_{\text{id}}) = f$.

(2) To show transitivity in a univariate setup, we need to show that given any $f_1, f_2 \in \mathcal{F}$, 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 $F_1^{-1} \circ F_2$. Since f_1 is strictly positive, F_1^{-1} is well defined and γ is uniquely specified. Furthermore, since f_2 is strictly positive, we have $\dot{\gamma} > 0$ and $\gamma \in \Gamma$. □

This result implies that together the pair (f_p, γ) spans the full set $\mathcal{F}$, 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 Figure 4 where the inner disk denotes the set $\mathcal{F}_p$. The increasing rings around $\mathcal{F}_p$ represent the set $\{(f_p, \gamma) | f_p \in \mathcal{F}_p\}$, with γ belonging to progressively larger submanifolds of Γ. Please refer to [10] for more details.

5. Problem 2: Modal-constrained univariate density estimation. In this section, we focus on the problem of estimating univariate *pdf*s under arbitrary modality constraints. A similar geometric framework, albeit using frequentist approach, for modality constrained density estimation is presented in [9]. For simplicity, we will restrict to *pdf*s that satisfy the following conditions: It is strictly positive and continuous with an interval support and zero boundaries. (For further simplicity, we will assume that the support is $[0, 1]$.) Furthermore, we assume that the pdf has $m \geq 1$ well defined modes that lie in $(0, 1)$. Let f be such a *pdf* and suppose that the $2m + 1$ critical points of f are located at b_i, for $i = 0, \ldots, 2m$, with $b_0 = 0$ and $b_{2m} = 1$. Define the *height-ratio vector* of f to be $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{2m-2})$, where $\lambda_i = f(b_{i+1})/f(b_1)$ is the ratio of the height of the $(i+1)$st interior critical point to the height of the first (from the left) mode. We define $\mathcal{F}$ to be the set of all continuous densities on $[0, 1]$ with zero boundaries. Let

$\mathcal{F}_m \subset \mathcal{F}$ be the subset with m modes, and let $\mathcal{F}_{m,\lambda} \subset \mathcal{F}_m$ be a further subset of *pdf*s with height-ratio vector equal to λ.

PROPOSITION 5.1. The group Γ acts on the set $\mathcal{F}_{m,\lambda}$ by the mapping $\mathcal{F}_{m,\lambda} \times \Gamma \to \mathcal{F}_{m,\lambda}$, given by $(f * \gamma) = \frac{f \circ \gamma}{\int (f \circ \gamma)\ dt}$. This was referred to as shape and area-preserving action in Section 2. Furthermore, this action is transitive. That is, for any $f_1, f_2 \in \mathcal{F}_{m,\lambda}$, there exists a unique $\gamma \in \Gamma$ such that $f_2 = (f_1 * \gamma)$.

Proof. The new function $\tilde{f} \equiv (f, \gamma)$ is called the *time-warped* density or just *warped* density. To prove this theorem, we first have to establish that the warped density $\tilde{f}$ is indeed in the set $\mathcal{F}_{m,\lambda}$. Note that time warping by Γ and the subsequent global scaling do not change the number of modes of f since $\dot{\gamma}$ is strictly positive (by definition). The modes simply move to their new locations $\{\tilde{b}_i = \gamma^{-1}(b_i)\}$. Secondly, the height-ratio vector of $\tilde{f}$ remains the same as that of f. This is due to the fact that $\tilde{f}(\tilde{b}_i) \propto f(\gamma(\gamma^{-1}(b_i))) = f(b_i)$ and $\tilde{\lambda} = \tilde{f}(\tilde{b}_{i+1})/\tilde{f}(\tilde{b}_1) = f(b_{i+1})/f(b_1) = \lambda$. Next, we prove the compatibility property that for every $\gamma_1, \gamma_2 \in \Gamma$ and f, we have $(f * (\gamma_1 \circ \gamma_2)) = ((f * \gamma_1) * \gamma_2)$. Since

$$((f * \gamma_1) * \gamma_2) = \frac{\frac{f \circ \gamma_1}{\int (f \circ \gamma_1)\ ds} \circ \gamma_2}{\int (\frac{f \circ \gamma_1}{\int (f \circ \gamma_1)\ ds} \circ \gamma_2)\ dt} = \frac{f \circ (\gamma_1 \circ \gamma_2)}{\int (f \circ (\gamma_1 \circ \gamma_2))\ dt} = (f * (\gamma_1 \circ \gamma_2)) \ ,$$

this property holds.

Finally, we prove the transitivity property: given $f, \tilde{f} \in \mathcal{F}_{m,\lambda}$, there exists a unique $\gamma_0 \in \Gamma$ such that $\tilde{f} = (f * \gamma_0)$. Let h_f be the height of the first mode of f, and let $h_{\tilde{f}}$ be the height of the first mode of $\tilde{f}$. Then, define two nonnegative functions according to $g = f/h_f$ and $\tilde{g} = \tilde{f}/h_{\tilde{f}}$. Note that the height of both of their first modes is 1 and the height-ratio vector for the interior critical points is λ. Also, let the critical points of f and $\tilde{f}$ (and hence g and $\tilde{g}$, respectively) be located at b_i and $\tilde{b}_i$, respectively, for $i = 0, \ldots, 2m$. Since the modes are well defined, the function g is piecewise strictly-monotonous and continuous in the intervals $[b_t, b_{t+1}]$, for $t = 0, 1, \cdots, 2m-1$. Hence, within each interval $[g(b_t), g(b_{t+1})]$ there exists a continuous inverse of g, termed g_t^{-1}. Then, set $\gamma_1(x) = g_t^{-1}\big(\tilde{g}(x)\big), x \in [\tilde{b}_t, \tilde{b}_{t+1}]$ is such that $(g \circ \gamma_1) = \tilde{g}$, and hence $(f * \gamma_1) = \tilde{f}$. Note that the γ_1 is uniquely defined, continuous, increasing, but not differentiable at the finitely many critical points $\tilde{b}_i$ in general. Hence $\dot{\gamma}_1$ does not exist at those points. But $\dot{\gamma}_1$ can be replaced by a weak derivative of γ_1. Let D_γ be a weak derivative of γ_1 that is equal to $\dot{\gamma}_1$ wherever $\dot{\gamma}_1$ exists, and 1 otherwise. Define $\gamma_0 = \int D_\gamma$. Then γ_0 and γ_1 are equal and $\dot{\gamma}_0$ exists everywhere, and $(f * \gamma_0) = \tilde{f}$. □

Now note that $\mathcal{F}_m = \bigsqcup_\lambda \mathcal{F}_{m,\lambda}$. Thus, for $f_0 \in \mathcal{F}_m$, the estimation procedure entails (1) estimating the (unique) height ratio vector λ_0 such that $f_0 \in \mathcal{F}_{m,\lambda_0}$, (2) constructing an element $f_1 \in \mathcal{F}_{m,\lambda_0}$, and (3) estimating the optimal time-warping function γ_0 such that $f_0 = (f_1 * \gamma_0)$.

Finally, even though the theory for modal-constrained density-estimation was developed with the simplifying assumption that the densities are zero at the boundaries, the assumption can be dropped easily by considering the heights at the boundaries as extra parameters λ_{2m-1} and λ_{2m}.

6. Bayesian inference. We perform Bayesian inference to obtain a distribution for ξ given the observed values $X = (x_1, x_2, \ldots, x_n)$:

$$\pi(\xi|X) = \frac{L(X|\xi)\pi(\xi)}{P(X)}, \tag{6.1}$$

where X is the set of observations and ξ is the variable representing a univariate *pdf* being estimated. Here $L(X|\xi)$ is the likelihood function and $\pi(\xi)$ is the prior density for ξ. This posterior distribution of the parameters $\pi(\xi|X)$ can then be used to make inference on density of X. In what follows, we restrict ourselves to univariate density estimation, and leave the general case for future.

In the absence of any shape constraints, we have $\xi = c$, the coefficient vector for the tangent space representation of the warping functions. We take four simulated examples, and present the unconstrained Bayesian density estimates in Figure 5 . The underlying true densities in these four examples are as follows:

(1) $f_0 \sim ((1/3)\text{Beta}(1,3) + (1/3)\text{Beta}(1,4) + (1/3)\text{Beta}(3,15))$,
(2) $f_0 \sim (0.5\exp(3) + 0.5\mathcal{N}(1, 0.25))$ truncated to $[0,1]$,
(3) $f_0 \sim ((1/3)\text{Beta}(1,3) + (1/3)\text{Beta}(1,4) + (1/3)\text{Beta}(20,3))$, and
(4) $f_0 \sim ((1/3)\mathcal{N}(0.2, 0.1) + (1/3)\mathcal{N}(0.6, 0.05) + (1/3)\mathcal{N}(0.8, 0.1))$ truncated to $[0,1]$.

To obtain the density estimates, we assume an independent, mean-zero, Gaussian prior with fixed variance for each of the coefficients, and a Gaussian proposal density centered at the current MCMC state with standard deviation 0.1. Figure 5 shows a clear improvement in the shape of the estimate over the initial guess (taken to be $\mathcal{N}(0.5, 1)$, and truncated to $[0,1]$) in the interior of the support. However, the improvement near the boundaries are lacking. Also, we notice a wiggly structure in the estimates induced by the global nature of the Fourier basis elements.

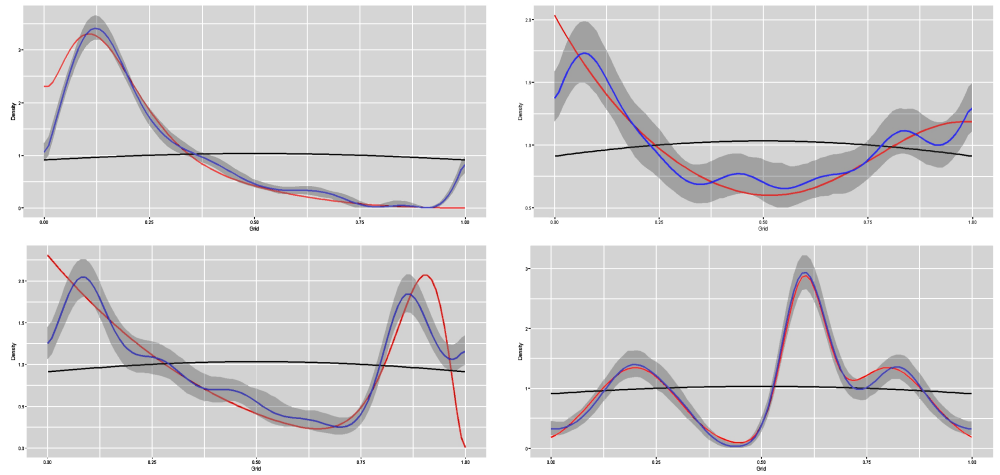

FIG. 5. True density shape (red); the initial guess (black); and the final unconstrained density estimate (blue) for the four simulated examples. The grey band represents the pointwise 95% posterior credible intervals.

Next we focus our attention to *modality-constrained* density estimation. From a model-based perspective, it is typical in the current literature to expand the function in a certain basis (B-spline, Bernstein polynomial, etc; see [41, 42, 44]) and transfer the constraints on the shape of the function to the vector of basis coefficients. We observe that this correspondence may not be one-to-one, limiting the approximation capability of the model. Further, in Bayesian implementations of such models, computation is hampered by the slow mixing of Markov chain Monte Carlo algorithms and there is a lack of theoretical support, especially concerning the uncertainty characterization from such models. In the presence of additional constraints on the parameter space, Markov chain Monte Carlo algorithms perform poorly. We develop a novel approach of incorporating a smooth version of the constraints, which facilitates mixing and convergence of the MCMC.

Note that we assume this density has the compact interval $[0, 1]$ as its support, is continuous, and has m modes in $(0, 1)$, but is not restricted to be zero at the boundaries. In this setup, the parameter vector ξ involves both the coefficient vector c and the height ratio vector λ. Since the posterior distribution of the parameters ξ is not conjugate due to the complicated form of the likelihood, we resort to MCMC techniques to sample from $\pi(\xi|X)$. We first discuss a novel technique to sample from nonconjugate distribution with both equality and inequality constraints on m.

6.1. *Sampling from nonconjugate distributions with linear inequality constraints.* We shall operate in a Bayesian framework which is attractive due to its natural ability to characterize uncertainty of estimation. A Bayesian specification of the above problem necessitates a prior distribution on ξ supported on $\mathcal{C}$, where $\mathcal{C}(\xi)$ is given by $\{\xi \in \mathbb{R}^p : A\xi \leq B\}$. Here A is an $l \times p$ matrix and B is a fixed p-dimensional vector with $p = J+2m$, where J is the number of basis elements used for tangent space representation of the warping functions, m is the number of modes of the true density, and l is the number of constraints in the model. A natural candidate is a truncated Gaussian prior,

$$\pi(\xi) \propto N(\xi; 0, \Sigma)\, \mathbb{1}_{\mathcal{C}}(\xi).$$

However, due to the complicated likelihood, the posterior distribution is not a truncated multivariate Gaussian. The presence of inequality constraints makes the problem worse. In the following, we describe a technique to sample from an intractable posterior subject to equality and inequality constraints on m. Consider the following posterior distribution:

$$\pi(\xi \mid X) \propto L(X \mid \xi)\mathcal{N}(\xi; 0, \Sigma)\, \mathbb{1}_{\mathcal{C}}(\xi).$$

We propose to approximate the indicator function using

$$\mathbb{1}(A\xi \leq B) \approx \prod_{i=1}^{l} \frac{\exp\{-\eta(a_i'\xi - B_i)\}}{1 + \exp\{-\eta(a_i'\xi - B_i)\}}, \tag{6.2}$$

where a_i is the ith row of A. Refer to [36] for a justification for such an approximation. One can then use a Metropolis Hastings algorithm to sample from the in-tractable posterior distribution. Next we describe the algorithm in detail. Let $c = (c_1, c_2, \ldots, c_J)$ be the

coefficient vector corresponding to the tangent space representation of the warping functions. Let $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{2m})$ be the height ratio vector. Let $\psi = (\psi_1, \psi_2, \ldots, \psi_{2m})$ be such that $\psi_i = \log \lambda_i, i = 1, 2, \ldots, 2m$. Then our parameter vector $\xi = (c, \psi) \in \mathbb{R}^{J+m}$.

We know that λ_1 is the first interior antimode from the left, λ_2 is the second mode from the left, and so on. λ_{2m-1} is the right boundary, an antimode, and λ_{2m} is the left boundary, another antimode. So the constraints on the height ratio parameters are as follows:

$$\left\{\begin{array}{lcl} \lambda_1 & < & 1 \\ \lambda_1 - \lambda_2 & < & 0 \\ \lambda_3 - \lambda_2 & < & 0 \\ & \cdots & \\ \lambda_{2m-1} - \lambda_{2m-2} & < & 0 \\ \lambda_{2m} & < & 1 \end{array}\right\}.$$

Reparameterizing in terms of ψ, we have

$$\left\{\begin{array}{lcl} \psi_1 & < & 0 \\ \psi_1 - \psi_2 & < & 0 \\ \psi_3 - \psi_2 & < & 0 \\ & \cdots & \\ \psi_{2m-1} - \psi_{2m-2} & < & 0 \\ \psi_{2m} & < & 0 \end{array}\right\}.$$

The parameter vector is $\xi = (c, \psi)$. For $m = 1$, define a $2 \times (J + 2)$ matrix A as $A = [0_{2\times J}\ I_2]$, where I_2 is the identity matrix of order 2. For $m > 1$, define a $2m \times (J+2m)$ matrix A as follows:

$$\left[\begin{array}{cccccccc} 0_{1\times J} & 1 & 0 & & \cdots & & & 0 \\ 0_{1\times J} & 1 & -1 & 0 & \cdots & & & 0 \\ 0_{1\times J} & 0 & -1 & 1 & 0 & \cdots & & 0 \\ 0_{1\times J} & 0 & 0 & 1 & -1 & \cdots & & 0 \\ & & & \cdots & & & & \\ 0_{1\times J} & 0 & \cdots & 0 & -1 & 1 & & 0 \\ 0_{1\times J} & 0 & & \cdots & & 0 & & 1 \end{array}\right].$$

Define a $2m \times 1$ vector B with $B_i = -\epsilon, i = 1, \ldots, 2m$. ϵ can be 10^{-6}, or any very small positive number. Then the constraint becomes $\mathbb{I}_{A\xi \le B}$.

Given a parameter vector $\xi = (c, \psi)$ satisfying $A\xi \le B$, the initial template function g and the warping function γ can be constructed as follows: Define $a_j = j/2m, j = 0, 1, \ldots, 2m$. Define $\lambda_i = \exp(\psi_i)$ and let $\Omega = \{\lambda_{2m}, 1, \lambda_1, \ldots, \lambda_{2m-1}\}$. Then set $g(a_j) = \Omega_j, j = 0, 1, \ldots, 2m + 1$. Obtain g for the other points in $[0, 1]$ through linear or polynomial interpolation. Take $\gamma = H(c)$, where H is the composite function defined in (3.3). Then the likelihood function $L(X|\xi)$ can be obtained as $L(X|\xi) = \prod_{i=1}^{n} g(\gamma(x_i))/[\int_0^1 g(\gamma(t))\, dt]$.

6.2. *Constraint on the upper bound of the number of modes.* The proposed framework allows a natural extension to have a more general shape constraint. Note that, for a fixed number M of peaks in the initial template, the deformed shape cannot introduce new peaks, and thus the final density shape has at most M peaks. The constraints on the height ratio parameters allow the density estimate to have *exactly* M modes. However, Proposition 6.1 shows that if the inequalities are relaxed, the density estimate can have any m number of modes as long as $m \leq M$.

PROPOSITION 6.1. Let g_λ be the template function with $\lambda \in \mathbb{R}^{2M}$ for some fixed M, and antimodes at the boundaries. Then g has $m \leq M$ modes.

Proof. Suppose that the locations of the $2M+1$ candidate critical points of the template function g be $a_i, i = 0, \ldots, 2M$. Note that because g is constructed via (linear) interpolation, it cannot have new critical points in (a_i, a_{i+1}). Thus, combined with the assumption that the boundary values are local minimas, the template function can have at most M local maximas and $M+1$ local minimas. Note that this situation occurs only when g alternates between monotonically increasing and monotonically decreasing in the intervals $[a_i, a_{i+1}]$. For example,if g is monotonic in the interval $[a_{i_0-1}, a_{i_0+1}]$ for some fixed i_0, then a_{i_0} is not a critical point anymore. Thus, g can have $m < M$ modes. □

Often in practice, it is natural to have an upper bound on the number of modes rather than an exact specification of the number of modes. This approach also serves as a technique of model selection among competing models corresponding to different modal constraints (unimodal versus bimodal, for example). This is because the relaxation of constraints allows optimization over all m-modal density estimates such that $m \leq M$ and returns the most likely estimate among all these models. Thus the number of modes of the final density estimate can be used to infer the number of modes in the true density that is "most likely" given the data.

6.3. *Simulation study.* For the algorithm we use a Markov Chain Monte Carlo (MCMC) technique to generate the posterior samples. We consider two prior structures for the parameters.

(1) First, the prior distribution of ξ was taken as $\mathcal{N}(0, 5I_{J+2m})\mathbb{I}_{A\xi \leq B}$. The constraints result in a nondifferentiable structure in the posterior distribution of the parameters. As before, $\mathbb{I}_{A\xi \leq B}$ is approximated by the differentiable function

$$\mathbb{I}_{A\xi \leq B} \approx \prod_{i=1}^{2m} \frac{\exp[-1000(a_i'\xi - B_i)]}{(1 + \exp[-1000(a_i'\xi - B_i)])},$$

where a_i is the ith row of A.

(2) Secondly, we propose a shrinkage prior setup as follows:

$$\pi(\xi) \propto \mathcal{N}(\xi; 0, \Sigma)\, \mathbb{I}_{A\xi \leq B}.$$

Here Σ is a diagonal matrix with $\Sigma_{ii} = s_i, i = 1, 2, \ldots, J + 2M$, with a prior distribution $s_i \sim \exp(1)$. The motivation of using an exponential prior for local scales s_i specific to each i is to allow large coordinate specific deviations, while shrinking the remaining s_i's close to 0. The indicator function is approximated by the differentiable function as before.

Denote by $(s_1, \ldots, s_{J+2m})'$ the vector s. To sample from the joint posterior distribution of ξ and s, given by $\pi(\xi, s \mid X)$, we resort to the Gibbs sampling algorithm and sample from the following full conditionals $\pi(\xi \mid s, X)$ and $\pi(s \mid \xi, X)$ instead. Sampling from $\pi(\xi \mid s, X)$ can be achieved using the Metropolis–Hastings scheme described before. To sample from $\pi(s \mid \xi, X)$, note that $\pi(s \mid \xi, X) = \prod_{i=1}^{J+2m} \pi(s_i \mid \xi_i)$. The conditional posterior distribution of the hyperpriors s_i is a generalized-inverse Gaussian (GIG) distribution given as

$$\pi(s_i \mid \xi_i) \sim \exp\{-(s_i + \xi_i^2/s_i)\}.$$

We use the `R` package `ghyp` to generate samples from the GIG distribution.

For illustrations we generate 1000 observations from six densities. We choose a Gaussian proposal density with σ as the standard deviation. For densities with sharper features, a smaller σ is required to capture the small changes in the coefficients of the higher order basis elements. However, a smaller σ forces the algorithm to take more steps to converge to the correct coefficients for the lower order basis elements. We employ both the Gaussian prior and the shrinkage prior setup for the parameters and compare the performances. Convergence was monitored using standard tests and diagnostic trace plots.

(1) $f_0 \propto \mathcal{N}(0.5, 0.1)$, a symmetric unimodal density truncated to the unit interval. Here, $\sigma = 0.001$ was chosen to capture the sharp peak of the true density. The MCMC algorithm was run for $400,000$ iterations with the first $300,000$ discarded as burn in. Also, up to 14 basis elements were used for the tangent space representation of the warping function.

(2) $f_0 \sim (1/3)\text{Beta}(1,3) + (1/3)\text{Beta}(1,4) + (1/3)\text{Beta}(3,15)$, a skewed unimodal example. Here, $\sigma = 0.001$ was used. The algorithm converged with $400,000$ iterations with the first $300,000$ discarded as burn in. Up to 14 basis elements were used for the tangent space representation of warping functions.

(3) $f_0 \propto 0.75\mathcal{N}(0.3, 0.2) + 0.25\mathcal{N}(0.75, .125)$, a bimodal density truncated to the unit interval. Here, $\sigma = 0.1$ was chosen. The algorithm converged within $40,000$ iterations where the first $20,000$ were discarded as burn in. For this example, up to 10 basis elements were used to represent the warping function.

(4) $f_0 \propto 0.5\mathcal{N}(0.1, 0.05) + 0.5\mathcal{N}(0.9, .05)$, a bimodal density truncated to the unit interval with very separated modes. Here, $\sigma = 0.1$ is used, and the algorithm converged within $50,000$ iterations with the first $30,000$ discarded as burn in. Here, up to 14 basis elements were used for approximating the warping function.

(5) $f_0 \propto (1/6)\mathcal{N}(0.2, 0.1) + (1/6)\mathcal{N}(0.6, 0.05) + (2/3)\mathcal{N}(0.8, 0.15)$, a trimodal density truncated to the unit interval. Here, $\sigma = 0.1$ was chosen. The algorithm was run for $200,000$ iterations where the first $100,000$ were removed as burn in. For this example up to 14 basis elements were used to represent the warping function.

(6) $f_0 \propto (1/3)\mathcal{N}(0, 0.1) + (1/3)\mathcal{N}(0.3, 0.05) + (1/3)\mathcal{N}(0.8, 0.1)$, a trimodal density truncated to the unit interval, with a mode at the boundary, two modes close, and the third mode separated. For this example, we use $\sigma = 0.01$ and $600,000$ MCMC iterations with the first $500,000$ discarded as burn in. Also, up to 14 basis elements were used for the tangent space representation.

Figure 6 illustrates the performance of the algorithm using a Gaussian prior. The improvement of the final estimate (blue line) from the initial guess (black line) is apparent. The grey band represents the 95% posterior credible interval. The choice of the proposal density is important given the complicated nature of the search space.

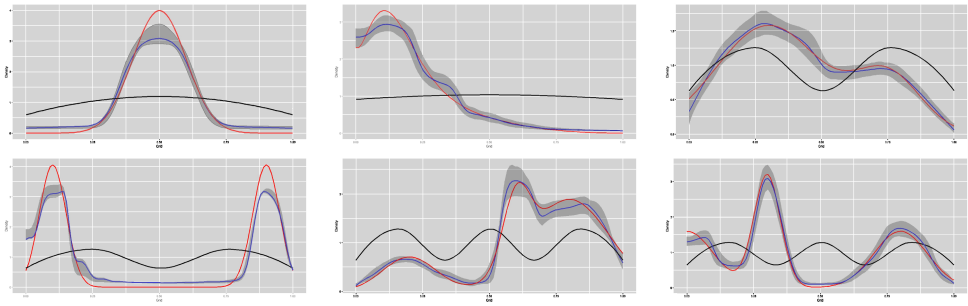

FIG. 6. An illustration of the true density shape (red); the initial guess (black); and the final density estimate (blue) for six simulated examples using Gaussian prior. The grey band represents the pointwise 95% posterior credible intervals.

To evaluate the average performance of the algorithm and the stability of the performance across samples, 50 samples each of sample size 1000 were chosen and the average performance of the estimate was recorded according to the $\mathbb{L}^2$ norm. As a benchmark for performance evaluation, a state-of-the-art kernel density estimate is used. The bandwidth for the kernel technique is chosen via the unbiased cross validation method. The kernel technique does not take into account the modality constraints, but since the sample size is fairly large, the estimate can be expected to have the correct number of modes most of the times. Also, as an unconstrained density estimator, the kernel technique is one of the most popular and widely used techniques and hence is a good benchmark for the performance of the proposed technique.

Table 1 compares the performance of the proposed Bayesian technique with the kernel density estimate. The comparison is *not meant* to demonstrate superior performance of

TABLE 1. A quantitative analysis of the performance of the Bayesian Estimate with two prior structures versus the unconstrained kernel estimate for simulated examples. The table presents the mean (Mean) and the standard deviation (std.dev) of the $\mathbb{L}^2$loss function for the 50 samples. Acceptance indicates the mean acceptance rate (standard deviation in brackets) of the algorithm across the samples.

Example:	Bayesian with Gaussian Prior			Bayesian Shrinkage Prior			Kernel density	
	Mean	std.dev	Acceptance	Mean	std.dev	Acceptance	Mean	std.dev
(1)	1.82	0.37	33.7%(2.1%)	1.89	0.31	33.8%(3.7%)	1.06	0.31
(2)	1.42	0.27	35.9%(2.7%)	1.52	0.23	36.9%(2.4 %)	1.67	0.51
(3)	0.85	0.23	4.3%(3%)	0.87	0.19	7.1% (2.6%)	0.83	0.21
(4)	3.47	1.2	0.04%(0.01%)	4.53	1.19	0.2%(0.07%)	1.47	0.33
(5)	1.26	0.32	0.3%(0.2%)	1.21	0.21	1.3% (0.3%)	1.14	0.28
(6)	2.94	0.99	0.3%(0.3%)	1.91	0.69	2.2%(0.4%)	1.58	0.69

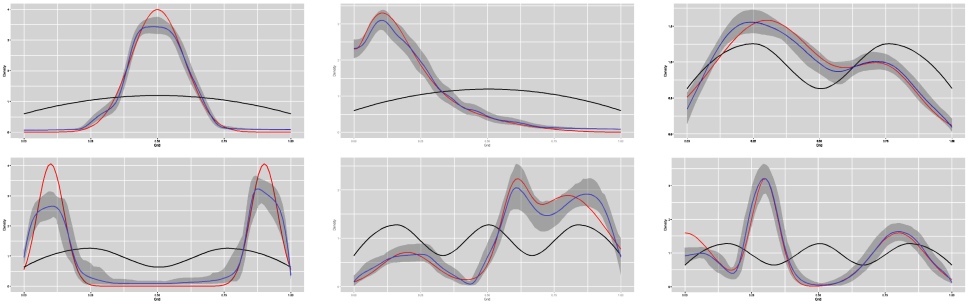

FIG. 7. An illustration of the true density shape (red); the initial guess (black); and the final density estimate (blue) for six simulated examples using shrinkage prior. The grey band represents the pointwise 95% posterior credible intervals.

our method compared to kernel methods. In fact, in terms of estimating the density itself in the $\mathbb{L}^2$ norm, it is expected that our approach will have inferior performance with respect to an unconstrained density estimator which is designed to approximate the unknown density arbitrarily well. On the other hand, our estimator with a prespecified number of modes avoids having spurious modes in the tails facilitating interpretation in a real-data scenario. We, however, notice that the performance of the proposed technique is quite close to that of the kernel technique for the bimodal and the trimodal examples. One caveat is that when the underlying density has a sharp peak (Examples (1) and (5)), this was not captured well. Since the acceptance rate of the algorithm is low (unless σ is very small), our approach fails to accurately capture the sharp peak (Examples (1) and (5)). The reason for the low acceptance rate is the lack of flexibility in the model to capture higher variability in certain coordinates and lower variability in others. The shrinkage priors offer more flexibility by allowing certain coordinates to deviate more, while shrinking the others towards zero. The shrinkage prior structure offers a higher acceptance rate and a faster convergence rate than the Gaussian prior counterpart in most of the examples. However, for both the prior structures, the average loss functions of the posterior mean density estimate are quite similar. The results are illustrated in Table 1. The middle panel of Table 1 indicates the clear improvement in acceptance rates, especially for the bimodal and trimodal cases. Also the standard deviations of the performances are much lower, indicating more stable estimates and faster convergence. The performance of the algorithm on the simulated examples is illustrated in Figure 7.

7. Application to electricity consumption. Here we illustrate an application of our approach to shape-constrained density estimation using electricity consumption data. This data was collected for households in Tallahassee, FL, at 30-minute intervals, on weekdays over a large period, and our goal is to model electricity consumption as a random variable and estimate its distribution (for a random household on a typical weekday). The electricity consumption is expected to follow a *bimodal* distribution corresponding to low and high electricity consumption. The low relates to times when the

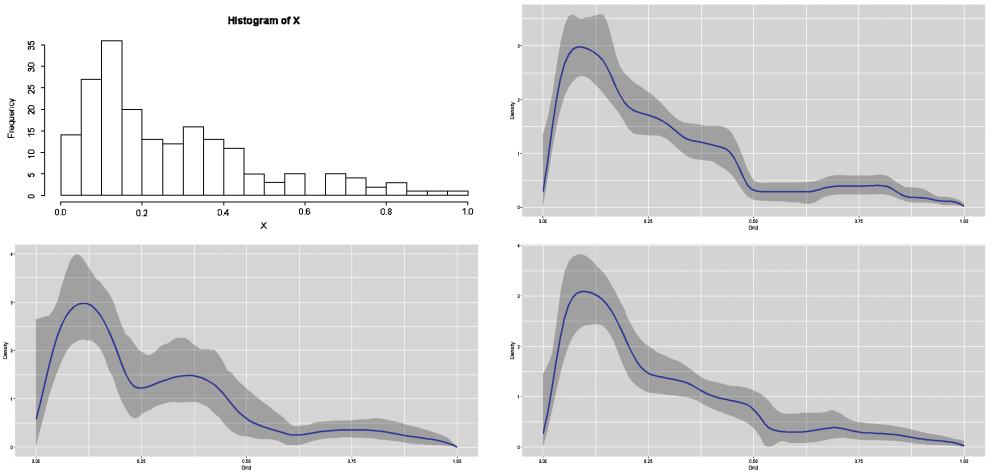

FIG. 8. The histogram (top-left); and density estimate (the remaining) for electricity consumption in a household in Tallahassee.

members of the households are not at home, or at home but not using any heavy appliances. High electricity consumption relates to times when high power appliances, such as the washer, heater, etc., are turned on.

When we study empirical distributions of electricity we observe three modes, perhaps corresponding to situations: (a) the residents are not at home, (b) the residents are home but not using heavy electricity consuming appliances, and (c) the residents are using heavy electricity consuming appliances. Figure 8 top-left panel shows a histogram for the electricity consumption and the remaining panels show the density estimates along with the pointwise 95% posterior credible intervals. The top-right panel presents the density estimate corresponding to the bimodal ($M = 2$ fixed) model, and the bottom-left panel presents the density estimate corresponding to the *trimodal* ($M = 3$ fixed) model. Note that the bimodal model replaces the left two modes of the trimodal model with a single mode with a long right tail. Finally, when we put an upper bound on the number of modes ($M \leq 3$), we obtain the estimate in the bottom right panel of Figure 8. Interestingly, this estimate supports the bimodal model, with a large dominant mode and a much smaller second mode on the tails.

8. Discussion. This paper introduces a novel Bayesian setup to perform density estimation with or without shape constraints using a *deformable template* approach. Here the *pdf*s are modeled using the actions of the deformation groups which, in turn, are estimated using the tangent space representations. The simulation studies show that the algorithm provides a significant improvement over initial estimates, and it is quite close to the benchmark kernel density estimate in practical performance. This framework results in a proper shape-constrained density estimator where shape implies a fixed number or an upper bound on the number of modes. This Bayesian setup is one of the first techniques to provide a density estimate $(\hat{f}|M)$, conditioned on the number of modes, and also provides

a measure of uncertainty measurement through posterior credible intervals. This setup allows for devising a probabilistic hypothesis testing algorithm on the number of modes present in the underlying distribution, although that is not explored here.

The algorithm is naturally computationally expensive since the goal here is to recover the posterior distribution of the shape as opposed to a point estimate. The algorithm performs a random walk without taking the constraint information into account and, thus, the acceptance rate is low unless the proposal density has a very low variance. If one tries a low variance proposal density, the samples become highly correlated. The acceptance rate is further reduced due to constraints. Finally, the performance of the algorithm in both of the prior setups was heavily dependent on the choice of σ in the proposal density.

With these observations, it seems natural to employ an algorithm that takes informed steps by utilizing information from the data. Thus, the Hamiltonian Monte Carlo (HMC) approach would be a more appropriate choice in this situation. HMC avoids the random walk behavior and sensitivity to correlated parameters by taking a series of steps informed by first-order gradient information. Another possible approach to avoid the choice of σ is to use elliptical slice sampling.

A natural direction for future work is to extend these ideas in higher dimensions, because the group actions are still valid in multi-dimensional setups. However, the optimization over the diffeomorphism group in higher dimensions becomes challenging and is not discussed in this paper.

Acknowledgments. The authors would like to thank the City of Tallahassee for providing the electricity consumption data. The contents of this paper represent the authors' opinion and do not reflect the official view of the City of Tallahassee.

References

[1] Yali Amit, Ulf Grenander, and Mauro Piccioni, *Structural image restoration through deformable templates*, Journal of the American Statistical Association **86** (1991), no. 414, 376–387.

[2] Richard E Barlow, *Statistical inference under order restrictions; the theory and application of isotonic regression*, 1972.

[3] Pierre C. Bellec and Alexandre B. Tsybakov, *Sharp oracle bounds for monotone and convex regression through aggregation*, J. Mach. Learn. Res. **16** (2015), 1879–1892. MR3417801

[4] Anirban Bhattacharya, Debdeep Pati, and David B Dunson, *Latent factor density regression models* (2012).

[5] Peter J. Bickel and Jianqing Fan, *Some problems on the estimation of unimodal densities*, Statist. Sinica **6** (1996), no. 1, 23–45. MR1379047

[6] Lucien Birgé, *Estimation of unimodal densities without smoothness assumptions*, Ann. Statist. **25** (1997), no. 3, 970–981, DOI 10.1214/aos/1069362733. MR1447736

[7] Hugh D. Brunk, *Estimation of isotonic regression*, University of Missouri-Columbia, 1969.

[8] Lawrence J. Brunner and Albert Y. Lo, *Bayes methods for a symmetric unimodal density and its mode*, Ann. Statist. **17** (1989), no. 4, 1550–1566, DOI 10.1214/aos/1176347381. MR1026299

[9] Sutanoy Dasgupta, Debdeep Pati, Ian H. Jermyn, and Anuj Srivastava, *Shape-Constrained Univariate Density Estimation*, ArXiv e-prints (April 2018), available at `1804.01458`.

[10] Sutanoy Dasgupta, Debdeep Pati, and Anuj Srivastava, *A geometric framework for density modeling*, arXiv preprint arXiv:1701.05656 (2017).

[11] Hassan Doosti and Peter Hall, *Making a non-parametric density estimator more attractive, and more accurate, by data perturbation*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **78** (2016), no. 2, 445–462, DOI 10.1111/rssb.12120. MR3454204

[12] Michael D. Escobar and Mike West, *Bayesian density estimation and inference using mixtures*, J. Amer. Statist. Assoc. **90** (1995), no. 430, 577–588. MR1340510
[13] Fuchang Gao and Jon A. Wellner, *Global rates of convergence of the MLE for multivariate interval censoring*, Electron. J. Stat. **7** (2013), 364–380, DOI 10.1214/13-EJS777. MR3020425
[14] Ulf Grenander, *On the theory of mortality measurement. II*, Skand. Aktuarietidskr. **39** (1956), 125–153 (1957). MR0093415
[15] U. Grenander, Y. Chow, and D. M. Keenan, *Hands*: *A pattern-theoretic study of biological shapes*, Research Notes in Neural Computing, vol. 2, Springer-Verlag, New York, 1991. MR1084371
[16] Peter Hall and Li-Shan Huang, *Unimodal density estimation using kernel methods*, Statist. Sinica **12** (2002), no. 4, 965–990. MR1947056
[17] Peter Hall, Simon J. Sheather, M. C. Jones, and J. S. Marron, *On optimal data-based bandwidth selection in kernel density estimation*, Biometrika **78** (1991), no. 2, 263–269, DOI 10.1093/biomet/78.2.263. MR1131158
[18] Clifford Hildreth, *Point estimates of ordinates of concave functions*, J. Amer. Statist. Assoc. **49** (1954), 598–619. MR0065093
[19] Nils Lid Hjort and Ingrid K. Glad, *Nonparametric density estimation with a parametric start*, Ann. Statist. **23** (1995), no. 3, 882–904, DOI 10.1214/aos/1176324627. MR1345205
[20] Alan Julian Izenman, *Recent developments in nonparametric density estimation*, J. Amer. Statist. Assoc. **86** (1991), no. 413, 205–224. MR1137112
[21] Sonia Jain and Radford M. Neal, *A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model*, J. Comput. Graph. Statist. **13** (2004), no. 1, 158–182, DOI 10.1198/1061860043001. MR2044876
[22] Maria Kalli, Jim E. Griffin, and Stephen G. Walker, *Slice sampling mixture models*, Stat. Comput. **21** (2011), no. 1, 93–105, DOI 10.1007/s11222-009-9150-y. MR2746606
[23] Roger Koenker and Olga Geling, *Reappraising medfly longevity: a quantile regression survival analysis*, J. Amer. Statist. Assoc. **96** (2001), no. 454, 458–468, DOI 10.1198/016214501753168172. MR1939348
[24] S. Kundu and D. B. Dunson, *Latent factor models for density estimation*, Biometrika **101** (2014), no. 3, 641–654, DOI 10.1093/biomet/asu019. MR3254906
[25] Peter J. Lenk, *The logistic normal distribution for Bayesian, nonparametric, predictive densities*, J. Amer. Statist. Assoc. **83** (1988), no. 402, 509–516. MR971380
[26] Peter J. Lenk, *Towards a practicable Bayesian nonparametric density estimator*, Biometrika **78** (1991), no. 3, 531–543, DOI 10.1093/biomet/78.3.531. MR1130921
[27] Tom Leonard, *Density estimation, stochastic processes and prior information*, J. Roy. Statist. Soc. Ser. B **40** (1978), no. 2, 113–146. With discussion. MR517434
[28] Qi Li and Jeffrey Scott Racine, *Nonparametric econometrics*: *Theory and practice*, Princeton University Press, Princeton, NJ, 2007. MR2283034
[29] Steven N MacEachern and Peter Müller, *Estimating mixture of dirichlet process models*, Journal of Computational and Graphical Statistics **7** (1998), no. 2, 223–238.
[30] Mary C. Meyer, *An alternative unimodal density estimator with a consistent estimate of the mode*, Statist. Sinica **11** (2001), no. 4, 1159–1174. MR1867337
[31] Washington Mio, Anuj Srivastava, and Shantanu Joshi, *On shape of plane elastic curves*, International Journal of Computer Vision **73** (2007), no. 3, 307–324.
[32] Peter Müller, Alaattin Erkanli, and Mike West, *Bayesian curve fitting using multivariate normal mixtures*, Biometrika **83** (1996), no. 1, 67–79, DOI 10.1093/biomet/83.1.67. MR1399156
[33] B. L. S. Prakasa Rao, *Estimation of a unimodal density*, Sankhyā Ser. A **31** (1969), 23–36. MR0267677
[34] Murray Rosenblatt, *Remarks on some nonparametric estimates of a density function*, Ann. Math. Statist. **27** (1956), 832–837, DOI 10.1214/aoms/1177728190. MR0079873
[35] S. J. Sheather and M. C. Jones, *A reliable data-based bandwidth selection method for kernel density estimation*, J. Roy. Statist. Soc. Ser. B **53** (1991), no. 3, 683–690. MR1125725
[36] Allyson Souris, Anirban Bhattacharya, and Debdeep Pati, *The soft multivariate truncated normal distribution*, arXiv preprint arXiv:1807.09155 (2018).
[37] Anuj Srivastava, Eric Klassen, Shantanu H. Joshi, and Ian H. Jermyn, *Shape analysis of elastic curves in Euclidean spaces*, IEEE Trans. PAMI **33** (2011), 1415–1428.
[38] Anuj Srivastava and Eric P. Klassen, *Functional and shape data analysis*, Springer Series in Statistics, Springer-Verlag, New York, 2016. MR3821566

[39] Surya T. Tokdar, *Towards a faster implementation of density estimation with logistic Gaussian process priors*, J. Comput. Graph. Statist. **16** (2007), no. 3, 633–655, DOI 10.1198/106186007X210206. MR2351083

[40] Surya T. Tokdar, Yu M. Zhu, and Jayanta K. Ghosh, *Bayesian density regression with logistic Gaussian process and subspace projection*, Bayesian Anal. **5** (2010), no. 2, 319–344, DOI 10.1214/10-BA605. MR2719655

[41] Bradley C. Turnbull and Sujit K. Ghosh, *Unimodal density estimation using Bernstein polynomials*, Comput. Statist. Data Anal. **72** (2014), 13–29, DOI 10.1016/j.csda.2013.10.021. MR3139345

[42] J. Wang and S. K. Ghosh, *Shape restricted nonparametric regression with Bernstein polynomials*, Comput. Statist. Data Anal. **56** (2012), no. 9, 2729–2741, DOI 10.1016/j.csda.2012.02.018. MR2915158

[43] Edward J. Wegman, *Maximum likelihood estimation of a unimodal density. II*, Ann. Math. Statist. **41** (1970), 2169–2174, DOI 10.1214/aoms/1177696724. MR0267681

[44] Matthew W. Wheeler, David B. Dunson, Sudha P. Pandalai, Brent A. Baker, and Amy H. Herring, *Mechanistic hierarchical Gaussian processes*, J. Amer. Statist. Assoc. **109** (2014), no. 507, 894–904, DOI 10.1080/01621459.2014.899234. MR3265664

[45] Laurent Younes, *Computable elastic distances between shapes*, SIAM J. Appl. Math. **58** (1998), no. 2, 565–586, DOI 10.1137/S0036139995287685. MR1617630

[46] Laurent Younes, Peter W. Michor, Jayant Shah, and David Mumford, *A metric on shape space with explicit geodesics*, Atti Accad. Naz. Lincei Rend. Lincei Mat. Appl. **19** (2008), no. 1, 25–57, DOI 10.4171/RLM/506. MR2383560