Technometrics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Modality-Constrained Density Estimation via Deformable Templates

Sutanoy Dasgupta , Debdeep Pati , Ian H. Jermyn & Anuj Srivastava

To cite this article: Sutanoy Dasgupta , Debdeep Pati , Ian H. Jermyn & Anuj Srivastava (2021): Modality-Constrained Density Estimation via Deformable Templates, Technometrics, DOI: 10.1080/00401706.2020.1867647

To link to this article: https://doi.org/10.1080/00401706.2020.1867647

View supplementary material

Published online: 01 Feb 2021.

Submit your article to this journal

Article views: 65

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=utch20

TECHNOMETRICS
2021, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/00401706.2020.1867647

Taylor & Francis
Taylor & Francis Group

Check for updates

Modality-Constrained Density Estimation via Deformable Templates

Sutanoy Dasgupta[a], Debdeep Pati[b], Ian H. Jermyn[c], and Anuj Srivastava[a]

[a]Department of Statistics, Florida State University, Tallahassee, FL; [b]Department of Statistics, Texas A&M University, College Station, TX; [c]Department of Mathematics and Statistics, Durham University, Durham, UK

ABSTRACT

Estimation of a probability density function (pdf) from its samples, while satisfying certain shape constraints, is an important problem that lacks coverage in the literature. This article introduces a novel geometric, deformable template constrained density estimator (`dtcode`) for estimating pdfs constrained to have a given number of modes. Our approach explores the space of thus-constrained pdfs using the set of shape-preserving transformations: an arbitrary template from the given shape class is transformed via a shape-preserving transformation to obtain the final optimal estimate. The search for this optimal transformation, under the maximum-likelihood criterion, is performed by mapping transformations to the tangent space of a Hilbert sphere, where they are effectively linearized, and can be expressed using an orthogonal basis. This framework is first applied to (univariate) unconditional densities and then extended to conditional densities. We provide asymptotic convergence rates for `dtcode`, and an application of the framework to the speed distributions for different traffic flows on Californian highways.

ARTICLE HISTORY
Received September 2019
Accepted November 2020

KEYWORDS
Conditional densities; Deformation group; Density estimation; Modality constraints; Shape constraints; Sieve estimation

1. Introduction

The estimation of a probability density function (pdf) from its samples is a fundamental problem in statistics, with a multitude of applications in different fields. A subproblem, involving estimation of a pdf given some prior knowledge about the shape of this pdf, is also an important problem. In practice, the prior knowledge stems from a scientific understanding of the underlying process. It is therefore important that the estimate be consistent with the prior shape knowledge for it to be interpretable and practically useful as an analytical tool. While a great deal of past research has gone into *shape-constrained* density estimation, these articles have dealt with very specific shape constraints, including log-concavity, monotonicity, and unimodality; there is little to no literature on optimization-based estimation of pdfs with multimodal shape constraints.

The earliest estimate for a unimodal density was given by Grenander (1956), who showed that a particular, natural class of estimators for unimodal densities is not consistent, and presented a modification that is consistent. Over the last several decades, a large number of articles have been written analyzing the properties of the *Grenander estimator* (e.g., Rao 1969; Izenman 1991) and its modifications (Birge 1997). An estimator using a maximum likelihood approach was developed by Wegman (1970). The earlier articles assumed knowledge of the position and value of the mode, and applied monotonic estimators over subintervals on either side of it; later articles (e.g., Meyer 2001; Bickel and Fan 1996) include an additional mode-estimation step. Bayesian methods have also been developed (Brunner and Lo 1989). Turnbull and Ghosh (2014), in addition to describing an estimator that uses Bernstein polynomials with the weights chosen to satisfy the unimodality constraint, also provided a useful summary of recent results on unimodal density estimation.

The obvious extension to multimodality constraints is of great practical importance because multimodal densities occur abundantly in nature; in particular, many biological processes are expected to show a known multimodal structure. For example, the DNA methylation profile in humans shows a bimodal structure corresponding to hypomethylated and hypermethylated regions: see Harris et al. (2010) and references therein; while the rate of nucleotide substitutions in DNA sequence (in non-CG-nucleotides) shows a trimodal density corresponding to accelerated, conserved, and neutral substitution rates: see Pollard et al. (2010) and references therein. In industrial and electrical engineering, household electricity consumption patterns and traffic patterns have been known to follow multimodal distributions.

1.1. Challenges and Current Literature

The important challenges in shape-constrained estimation are to characterize the set of all density functions satisfying the desired shape constraints, and to solve the maximum likelihood estimation problem on that space. Shape-constrained estimation problems would seem to encourage a geometric approach, but the use of geometry in density estimation has in fact been sparse in the literature: to the best of our knowledge, there is *no current method that can impose a multimodality constraint on an estimated density and provide optimality in some way*. However, multimodality constraints have been studied in the case of function estimation: for example, see the very recent

CONTACT Sutanoy Dasgupta s.dasgupta@stat.fsu.edu Department of Statistics, Florida State University, Tallahassee, FL 32304.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TECH.
© 2021 American Statistical Association and the American Society for Quality

article by Wheeler, Dunson, and Herring (2017) and references therein. Here we summarize the literature that is most relevant to the problem of density estimation under shape constraints.

Hall and Huang (2002) introduced a tilting approach to convert an unconstrained density to an estimate within a *unimodal shape class*. However, the resultant density estimate often directly contradicts the available data by having zero likelihood even at the data points themselves, and is thus not appropriate as an exploratory tool. Another article (Cheng, Gasser, and Hall 1999) proposes to start with a template unimodal density and provide a sequence of transformations that when applied to the template both keep the result unimodal, and "improve" the estimate in some sense. However, the method is ad hoc, and asymptotic convergence of the estimates, although seen empirically, is not guaranteed. Very recently, Wolters and Braun (2018) introduced a technique that solves the limitations of the approach in Hall and Huang (2002). Specifically, they provide an algorithm to find a constrained estimate that is the *nearest* to an unconstrained kernel density estimate (under the integrated squared error loss function), and that can handle up to *bimodal* constraints. However, this method provides an estimate that satisfies the shape constraint only on a prespecified grid in the support, so that the estimate need not lie in the correct shape class, in principle. Since their construction of the constrained estimate involves smoothing out the spurious peaks of the initial unconstrained estimate, the resultant shape contains spurious flat spots, which once again limits the interpretability of the estimate. Finally, this estimate is not designed to be optimal under any specific criterion. This issue is also present in kernel density estimators, where one can always choose a bandwidth to ensure a given number of modes, but the resulting density is not optimal in any sense for a finite sample size.

Recently, Dasgupta, Pati, and Srivastava (2021) introduced a geometric approach for exploring the space of *all* probability densities to perform unconstrained density estimation. In this approach, one starts with an efficient initial estimate, perhaps from a parametric family, and then transforms it into the desired optimal density using elements of a diffeomorphism group. The problem therefore shifts to finding the optimal transformation under the chosen criterion. However, no shape constraints are imposed on the estimated density.

1.2. Proposed Formulation and Its Novelty

In the current article, we take a principled and geometrically intuitive maximum-likelihood approach to the problem of modality-constrained density estimation. The primary contribution of this article is to construct a framework that can handle any general modality constraint, and can provide smooth interpretable maximum likelihood estimates within a specified shape class.

For this purpose, we develop a novel modification of the geometric approach used by Dasgupta, Pati, and Srivastava (2021). The method starts with a *template* density from the desired shape class, and then deforms it into the optimal estimate from that shape class. We shall call this estimator deformable template constrained density estimator or `dtcode`. The advantages of `dtcode` over existing methods are as follows.

First, while estimation is based on deformation or transformation of an initial template as in Cheng, Gasser, and Hall (1999), we apply only a single transformation rather than a possibly nonconvergent sequence. Coupled with a small number of other parameters, this transformation constitutes a parameterization of the whole of the shape class of interest.

Second, we use a broader notion of shape than previous work: in its simplest form, we constrain the pdf to possess a fixed, but arbitrary, number of modes; we also consider more general cases in the supplementary materials. The shape constraint is fully captured in the initial template itself. As a result, the subsequent estimation of the transformation is independent of the constraint, providing much greater stability in practical performance with respect to higher modality constraints than methods such as Wolters and Braun (2018).

Third, we use (penalized) maximum likelihood estimation, which guarantees optimality in principle, and allows the derivation of asymptotic rates of convergence to the true density.

The main difference between the current approach and Dasgupta, Pati, and Srivastava (2021) is in the choice of transformations used. Dasgupta, Pati, and Srivastava (2021) wish to parameterize the set of all positive densities. As a result they choose a set of transformations that act transitively, that is, any positive density may be transformed into any other. The necessary transformations take the standard form for a change of variable: a density is transformed by a warping of its domain: $p \mapsto (p \circ \gamma)\dot{\gamma}$, where p is positive probability density and the warping function γ is a diffeomorphism of the domain, that is, a one-to-one, differentiable map whose inverse is also differentiable. Here, $\dot{\gamma}$ is the derivative of γ, that is, $\dot{\gamma}(t) = \frac{\mathrm{d}\gamma(t)}{\mathrm{d}t}$ for all t in the domain of the diffeomorphism.

Clearly, these transformations are not suitable for our case because transitivity is not compatible with preserving the shape of a density, merely its normalization. We therefore propose a different set of transformations, which preserve both normalization and shape: they take the form $p \mapsto (p \circ \gamma)/\int (p \circ \gamma)dt$. The denominator renormalizes the density after the transformation in the numerator; together they preserve the shape of p, in a sense that we will now explain.

1.3. Overview of the Approach

A precise formulation of the problem is as follows: given independent samples $X = \{x_i\}, i = 1, \ldots, n$, from a pdf p_0, with a known number $M > 0$ of well-defined modes, estimate this density ensuring the presence of M modes in the solution. To do this, we construct a parameterization of the set of continuous densities with M modes, $\mathcal{P}_M$, as follows.

- Let the set of densities satisfying the shape constraint be denoted $\mathcal{P}_M = \{p : [0,1] \to \mathbb{R}_+ : p(0) = p(1) = 0, p \text{ has } M \text{ interior modes}\}$.
- Let the critical points of a pdf $p \in \mathcal{P}_M$ with M modes be located at $\{b_a : a \in \{0, \ldots, 2M\}\}$, with $b_0 = 0$ and $b_{2M} = 1$.
- We define the height ratio vector λ of p as the set of ratios of the height of the $(a+1)$th interior critical point to the height of the first (from the left) mode: $\lambda = (\lambda_1, \ldots, \lambda_{2M-2})$, where $\lambda_a = p(b_{a+1})/p(b_1)$. Please look at the top left panel

of Figure 2 for an illustration. The height ratio vector for the density p_0 illustrated here is simply $\lambda = (h_2/h_1, h_3/h_1)$.
- Let the subset of $\mathcal{P}_M$ with height ratio vector λ be denoted $\mathcal{P}_{M,\lambda}$. Note that the space $\mathcal{P}_M$ is the union $\cup_{\lambda} \mathcal{P}_{M,\lambda}$ of the individual spaces $\mathcal{P}_{M,\lambda}$ with different values of λ.

We then parameterize an arbitrary member of $\mathcal{P}_M$ by:

1. a height ratio vector $\lambda \in \Lambda_M$, where Λ_M is the set of all such vectors;
2. a diffeomorphism $\gamma \in \Gamma$, where $\Gamma = \{\gamma : [0,1] \to [0,1] : \dot{\gamma} > 0, \gamma(0) = 0, \gamma(1) = 1\}$ is the group of diffeomorphisms of $[0,1]$. Notably, the set Γ is a group, that is, it is closed under composition, has an identity element $\gamma_{id}(t) = t$, and each element γ has an inverse γ^{-1}.

The pdf represented by a pair λ and γ is then $p_{\lambda,\gamma} = (\tilde{p}_\lambda, \gamma) \in \mathcal{P}_{M,\lambda}$, where $\tilde{p}_\lambda \in \mathcal{P}_{M,\lambda}$ is an a priori fixed template function in $\mathcal{P}_{M,\lambda}$, and $(\cdot, \gamma)$ denotes the transformation of densities by elements of Γ mentioned earlier, which has the crucial property that it preserves λ.

Using this parameterization, we can construct the log-likelihood function

$$L(\lambda, \gamma | X) = \sum_i \log p_{\lambda,\gamma}(x_i), \tag{1}$$

and we can use maximum likelihood to estimate λ and γ.

We can generalize the method to a larger set of shape classes by defining a shape as a sequence of piecewise monotonically increasing, decreasing, and flat intervals that together constitute the entire density function. For example, an "N-shaped" density function is given by the sequence: *increasing-decreasing-increasing*. For any such sequence, we can construct a template density in the appropriate shape class, and proceed with estimation as before. The assumption $p_0(0) = p_0(1) = 0$ can also be relaxed, by considering the height ratios of the two boundaries as two extra parameters λ_0 and λ_{2M+1}. We discuss these ideas in more detail in Section 5 of the supplementary materials and present some simulated examples.

2. Geometric Representation of Densities

In this section, we show that the above construction does indeed provide a parameterization of $\mathcal{P}_M$, by first showing that Γ is large enough to allow us to reach any element of $\mathcal{P}_{M,\lambda}$ starting from a template $\tilde{p}_\lambda \in \mathcal{P}_{M,\lambda}$, and then showing how to construct such a template for each height ratio vector $\lambda \in \Lambda_M$.

Theorem 1. The set of transformations of the set $\mathcal{P}_{M,\lambda}$ by the mapping $\mathcal{P}_{M,\lambda} \times \Gamma \to \mathcal{P}_{M,\lambda}$, given by $(p, \gamma) = \frac{p \circ \gamma}{\int (p \circ \gamma)\, dt}$ is a group action. Furthermore, this action is transitive and free. That is, for any $p, \tilde{p} \in \mathcal{P}_{M,\lambda}$, there exists a unique $\gamma \in \Gamma$ such that $p = (\tilde{p}, \gamma)$.

The proof of the theorem is in the supplementary materials. The theorem shows that given a template $\tilde{p}_\lambda \in \mathcal{P}_{M,\lambda}$, we can uniquely represent any other pdf p with the same height-ratio vector (i.e., also in $\mathcal{P}_{M,\lambda}$) as a transformation of the template, that is, as $p = (\tilde{p}_\lambda, \gamma)$. What is more, any pdf in $\mathcal{P}_{M,\lambda}$ can serve as a template; it can thus be chosen for convenience' sake.

Figure 1 illustrates the height-ratio-vector-preserving effect of the transformations by applying several elements of Γ to a

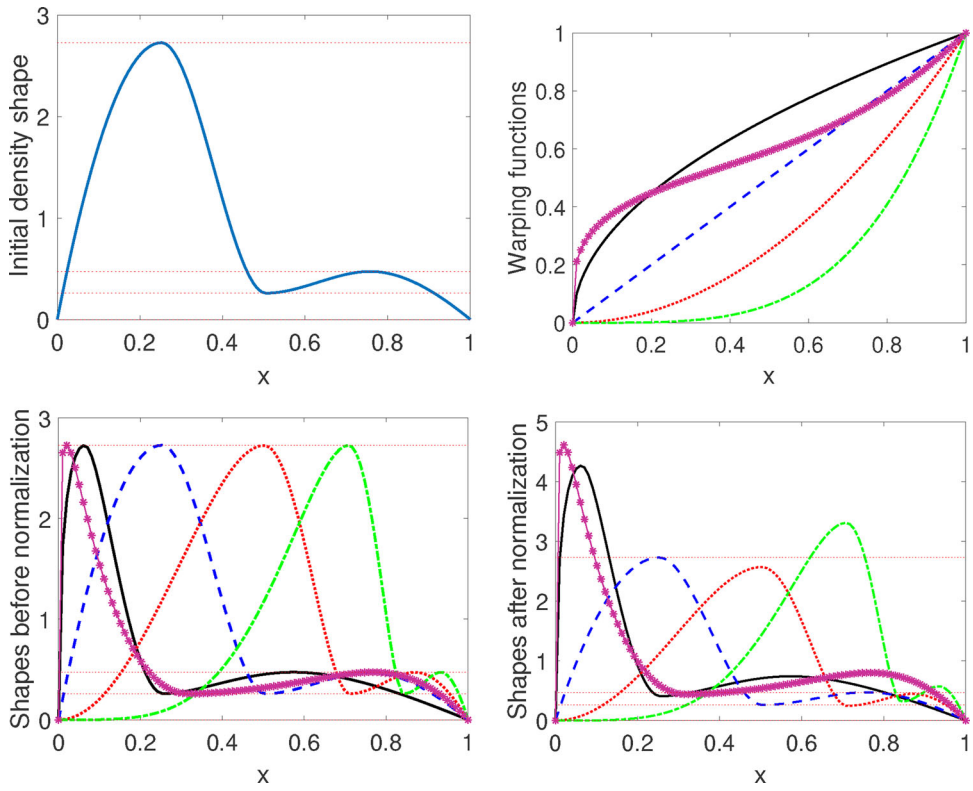

Figure 1. Top left: Initial density. Top right: Different warping functions. Bottom left: Shapes resulting from warping the initial density without renormalization. Bottom right: Resultant warped densities after renormalization.

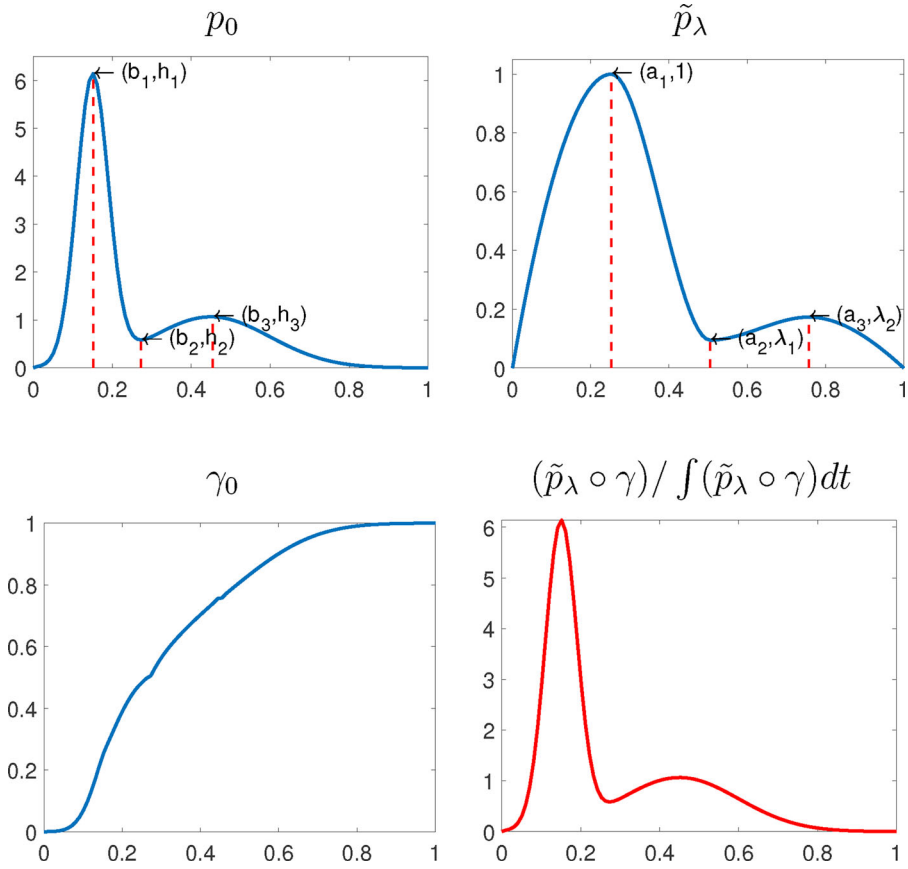

Figure 2. Top left: The original density. Top right: Initial template. Bottom left: The γ_0 transforming the template to original shape. Bottom right: Reconstructed density.

pdf in two stages. First, the numerator of the full transformation is shown (bottom-left); this stage preserves the heights of all extrema. Second, the pdf is renormalized by dividing by the denominator (bottom-right); this stage changes the heights, but still preserves the height-ratio vector.

How then do we construct a distinguished template element $\tilde{p}_\lambda \in \mathcal{P}_{M,\lambda}$? First we construct an unnormalized function g_λ with M modes and height ratio vector λ:

1. Set $g_\lambda(0) = g_\lambda(1) = 0$.
2. Divide the interval $[0, 1]$ into $2M$ equal intervals corresponding to the M modes and $M - 1$ interior antimodes, setting $a_j = j/2M$, $j \in [1, \ldots, (2M - 1)]$, the location of the jth critical point.
3. Set $g_\lambda(a_1) = 1$, and $g_\lambda(a_j) = \lambda_{j-1}$ for $j \in [2, \ldots, (2M - 1)]$.
4. The values of g_λ for all other points are obtained by linear interpolation.

We can now define $\tilde{p}_\lambda = g_\lambda/(\int g_\lambda) \in \mathcal{P}_{M,\lambda}$.

We have thus constructed a representation space $\Lambda_M \times \Gamma$, a set of coordinates for $\mathcal{P}_M$, where $\Lambda_1 = \{1\}$, and for $M > 1$, $\Lambda_M = \{\lambda \in \mathbb{R}_+^{(2M-2)} : \lambda_1 < 1, \lambda_1 < \lambda_2, \lambda_{2j+1} < \lambda_{2j}, \lambda_{2j+1} < \lambda_{2j+2}, j = 1, 2, \ldots, M - 2\}$, the conditions arising because the odd indices $\lambda_1, \lambda_3, \ldots, \lambda_{2M-3}$ correspond to antimodes, while the rest correspond to modes.

Figure 2 shows a simple example to illustrate this representation. The top left panel is a density that has $M = 2$ modes with critical points located at b_i and heights h_i. The top right panel shows the initial template function with $M = 2$ modes and critical points located at a_i and heights $\lambda_i = h_i/h_1$. The bottom left panel shows the warping function constructed according to the description in the proof of Theorem 1, while the last panel shows that using this warping function, we recover the original density.

So far, we have assumed that the densities are defined on $[0, 1]$. When the bounds of the density function are not known, they are estimated from the data X using the formula $A = \min(X) - \text{sd}(X)/\sqrt{n}$ and $B = \max(X) + \text{sd}(X)/\sqrt{n}$, where A and B are the lower and upper bounds, respectively, $\text{sd}(X)$ is the standard deviation of the observations, and n is the number of observations; these estimates are taken from Turnbull and Ghosh (2014). The data are then scaled to the unit interval, $z_i = (x_i - A)/(B - A)$, before proceeding with the rest of the estimation.

The framework readily extends to the situation where the true density has a general connected support $\mathcal{D}$ by generalizing from Γ to $\Gamma^* = \{\gamma : \mathcal{D} \to \mathcal{D}, \dot{\gamma} > 0, \gamma \text{ is boundary preserving}\}$. For example, if the support of the true density is the entire real line then we can set $\mathcal{D} = \mathbb{R} \cup \{\pm\infty\}$. However, from a practical standpoint, it is often beneficial to assume that the true density has compact rather than infinite support. Our experiments corroborate the findings in Wahba (1981), that it is preferable for the true density to have compact support and then to scale the data to the unit interval for density estimation. Thus, for the rest of the article, we always assume that $\mathcal{D} = [0, 1]$, and that the true density has its support on the unit interval.

3. Parameter Estimation

Having established a parameterization of the set of shape-constrained densities of interest, the next step is derive a procedure for estimating these parameters from data and specify the pdf estimator `dtcode`. We will use a maximum-likelihood framework, for which we must first specify the log-likelihood function and then solve the optimization problem for λ and γ. The optimization over Γ presents particular difficulties regardless of the likelihood function, and so we first describe how we deal with these.

3.1. Finite-Dimensional Representation of Warping Functions

In solving an optimization problem on Γ, we face two challenges. First, Γ is a nonlinear manifold, that is, it is not a vector space; and second, it is infinite-dimensional. We handle the nonlinearity by forming a map from Γ to a vector space. (This vector space happens to be the space tangent to the unit Hilbert sphere $\mathbb{S}_\infty$ as explained below.) We tackle infinite dimensionality by restricting to a finite-dimensional subspace of this vector space. Together, these two steps are equivalent to finding an increasing family of finite-dimensional subsets $\Gamma^J \subset \Gamma$ that can be flattened into vector spaces. This then allows us to represent any element $\gamma \in \Gamma^J$ using a finite orthogonal basis. Once we have a finite-dimensional representation of γ, we can optimize over this representation using standard techniques.

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, please refer to Chapter 4 of Srivastava and Klassen (2016)). Note that we can reconstruct γ from q_γ using $\gamma(t) = \int_0^t q_\gamma^2(s)\ ds$. In particular, since $||q_\gamma||^2 = \int_0^1 q_\gamma(t)^2 dt = \int_0^1 \dot\gamma(t) dt = \gamma(1) - \gamma(0) = 1$, we see that $q_\gamma \in \mathbb{S}_\infty$, where the unit Hilbert sphere $\mathbb{S}_\infty$ is defined by $\mathbb{S}_\infty \subset \mathbb{L}^2 = \{q : [0,1] \to \mathbb{R} : \int q^2(t)\, dt = 1\}$. We can also see that for any $q \in \mathbb{S}_\infty$, there is a γ_q that generates q given by $\gamma_q(t) = \int_0^t q^2(s)\ ds$.

The unit sphere $\mathbb{S}_\infty$ has known geometry Lang (2012), but is still not a vector space. However, it can easily be easily flattened into a vector space (locally) due to its constant curvature. A natural choice for this flattening is a bijective mapping, described next, to the vector space tangent to $\mathbb{S}_\infty$ at the point $\mathbf{1}$, a constant function with value 1. Note that $\mathbf{1}$ is the SRSF corresponding to $\gamma = \gamma_{\text{id}}(t) = t$, that is, the identity, making it a natural choice for the tangent space.) The tangent space of $\mathbb{S}_\infty$ at $\mathbf{1}$ is an infinite-dimensional vector space given by: $T_{\mathbf{1}}(\mathbb{S}_\infty) = \{v \in \mathbb{L}^2([0,1],\mathbb{R}) : \int_0^1 v(t)dt = \langle v, \mathbf{1}\rangle = 0\}$.

The bijective mapping between $\mathbb{S}_\infty$ and $T_{\mathbf{1}}(\mathbb{S}_\infty)$ is the so-called inverse exponential map:

$$\exp_{\mathbf{1}}^{-1}(q) : \mathbb{S}_\infty \to T_{\mathbf{1}}(\mathbb{S}_\infty)\ ,$$
$$v = \exp_{\mathbf{1}}^{-1}(q) = \frac{\theta}{\sin(\theta)}(q - \mathbf{1}\cos(\theta))\ , \tag{2}$$

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

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, since $\cos^{-1}(\langle \mathbf{1}, q\rangle) < \pi$, the norm $||v|| = \sqrt{\int_0^1 v(t)^2 dt} = \theta < \pi$ for any $v = \exp_{\mathbf{1}}^{-1}(q)$. Thus, the range of the inverse exponential map is not the entire $T_{\mathbf{1}}(\mathbb{S}_\infty)$, but a subset $V = \{v \in T_{\mathbf{1}}(\mathbb{S}_\infty) : ||v|| < \pi\}$.

To map points back from the tangent space to the Hilbert sphere, we reverse this process. This time we use the exponential map:

$$\exp_{\mathbf{1}}(v) : V \to \mathbb{S}_\infty\ , \quad \exp_{\mathbf{1}}(v) = \cos(||v||)\mathbf{1} + \frac{\sin(||v||)}{||v||}. \tag{3}$$

Finally, we can select any orthogonal basis $\mathcal{B} = \{b_j, j = 1,2,\dots\}$ of the Hilbert space $T_{\mathbf{1}}(\mathbb{S}_\infty)$ and express its elements v by their corresponding coefficients: $v(t) = \sum_{j=1}^\infty c_j b_j(t)$, where $c_j = \langle v, b_j\rangle$. The elements of such a basis are just functions in $\mathbb{L}^2([0,1],\mathbb{R})$ that are orthogonal to $\mathbf{1}$, that is, $\langle b_j, \mathbf{1}\rangle = 0$ for all j. One example is the Fourier basis excluding $\mathbf{1}$, but other bases, such as the cosine basis, splines, and Legendre polynomials, can also be used. Efromovich (2010) discussed different choices of basis functions and advocates the use of trigonometric bases for functions with compact support.

Given a basis $\mathcal{B} = \{b_j, j = 1,2,\dots\}$, one can define an infinite-dimensional space of coefficients $\mathcal{C} = \{c = (c_1, c_2, \dots) : \sum_{j=1}^\infty c_j b_j(t) \in V\}$. One can then truncate the basis expansion to approximate elements of V using finitely many coefficients. Suppose one uses J basis elements to approximate the tangent space elements. Then, the approximating space of coefficients will be denoted by $\mathcal{C}^J = \{c \in \mathbb{R}^J | \sum_{j=1}^J c_j b_j(t) \in V\}$. Note that $\mathcal{C}^J$ is a proper subset of $\mathbb{R}^J$ since it contains only 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 the transformation space Γ. We define a composite map $H : \mathcal{C}^J \to \Gamma$, as

$$\{c_j\} \in \mathcal{C}^J \xrightarrow{\{b_j\}} v = \sum_{j=1}^J c_j b_j \in V \xrightarrow{\exp_{\mathbf{1}}} q \in \mathbb{S}_\infty \to \gamma(t)$$
$$= \int_0^t q(s)^2 ds\ . \tag{4}$$

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 subset of Γ, denoted by Γ^J henceforth, and we pose the estimation problem on this subset. As J goes to infinity, this subset Γ^J converges to the full group Γ.

3.2. Joint Estimation of λ and γ

We use a joint maximum likelihood method to estimate the height ratios λ along with the coefficients corresponding to the estimate of γ. The maximum likelihood estimate of the underlying density, given the initial template function $\tilde p_\lambda$, is

$$\hat p(t) = \frac{\tilde p_{\hat\lambda}(\hat\gamma(t))}{\int_0^1 \tilde p_{\hat\lambda}(\hat\gamma(t))dt}\ , \quad t \in [0,1]\ , \tag{5}$$

where $\hat\gamma = H(\hat c)$, and

$$(\hat c, \hat\lambda) = \underset{c \in \mathcal{C}^J, \lambda \in \Lambda_M}{\text{argmax}} \left(\sum_{i=1}^n \left[\log\left(\tilde p_\lambda\left(\gamma_c(x_i)\right) / \int_0^1 \tilde p_\lambda\left(\gamma_c(t)\right)\ dt \right) \right] \right)\ . \tag{6}$$

Since this optimization is over a finite-dimensional Euclidean space, any numerical optimization package can be applied here. The objective function (6) is not convex, and so we use the Matlab function `fmincon` for optimization. However, `fmincon` can produce local solutions, and the `GlobalSearch` toolbox often yields better results, albeit at higher computational cost. The `GlobalSearch` toolbox is a multistarting algorithm that generates different trial points as initial values of the algorithm, and uses the trial point that converges to a local solution with the least objective function value. The algorithm and the method of generating these trial starting points are described in Ugray et al. (2007). Depending on the computational resource available, one can regulate the number of trial points generated, or can simply use the zero vector as a natural starting point.

The choice of J, the number of basis elements, is important. Too large J can result in overfitting and also put computational burden on the optimization algorithm which might get stuck in local, suboptimal solutions. We use a penalized version of the likelihood in (6), the standard Akaike's information criterion (AIC), to choose the optimal number of basis elements.

4. Simulation Study

For the numerical implementation of `dtcode`, we use the Fourier basis for the tangent space representation. We start with two basis elements, and increase the number up to a predecided limit; we then choose the result with the best AIC value. We chose AIC as the penalty on the number of basis elements because experiments suggested that BIC over-penalizes the number of parameters, causing the estimate to miss the sharper features of the true density. The code for `dtcode` is available online at *https://github.com/Sutanoy/Shapeconstrained_DensityEstimation*.

For illustration, we use sample sizes of 100, 500, and 1000. To evaluate the average performance, we generate 100 samples (of sample size 100, 500, and 1000, respectively) and evaluate the mean error and the standard deviation of the errors. For the error function, we considered the vector $\mathbb{L}^2$, $\mathbb{L}^1$, and $\mathbb{L}^\infty$ norms of the difference between the true density and the density estimate evaluated on 100 equidistant points across the support.

The average computational time for `dtcode` varies from around 20 sec for a sample of size 100, to 250 sec for a sample of size 1000, while optimizing over ten different possible parameter dimensions using 1000 trial points in the `GlobalSearch` algorithm, on an Intel(R) Core(TM) i7-3610QM CPU processor laptop.

4.1. Study 1

We start with two examples with one mode:

- Example 1: a symmetric unimodal pdf given by $p_0 = 0.8\,\mathcal{N}(0,4) + 0.2\,\mathcal{N}(0,0.5)$.
- Example 2: a unimodal pdf with contamination, given by $p_0 = 0.95\,\mathcal{N}(0,0.5) + 0.05\,\mathcal{N}(3,1)$.

In Figure 3, we use the $\mathbb{L}^2$ loss function values calculated for 100 samples of size 100 to compare the results obtained with `dtcode` (leftmost column) to those obtained using the `umd` package developed by Turnbull and Ghosh (2014) (center column) and the `scdensity` package introduced in Wolters and Braun (2018) (rightmost column). The upper and lower rows show examples 1 and 2. In each plot, the true density is shown

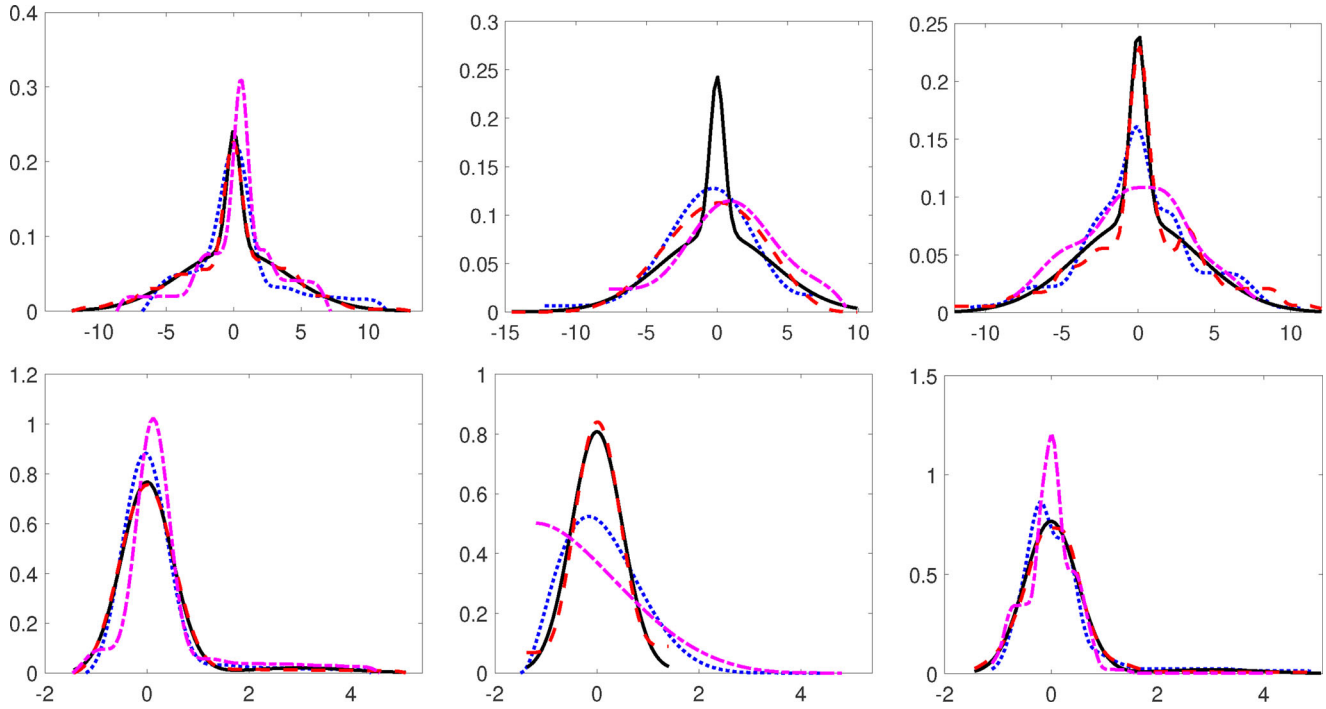

Figure 3. Performance of `dtcode` (left) versus `umd` (center), and `scdensity` (right) for examples 1 (top row) and 2 (bottom row) of Study 1, using the $\mathbb{L}^2$ loss function values calculated for 100 samples of size 100. The true density is shown as a solid line; the estimated density with best performance as a dashed line; with median performance as a dotted line; and with worst performance as a dashed-dotted line.

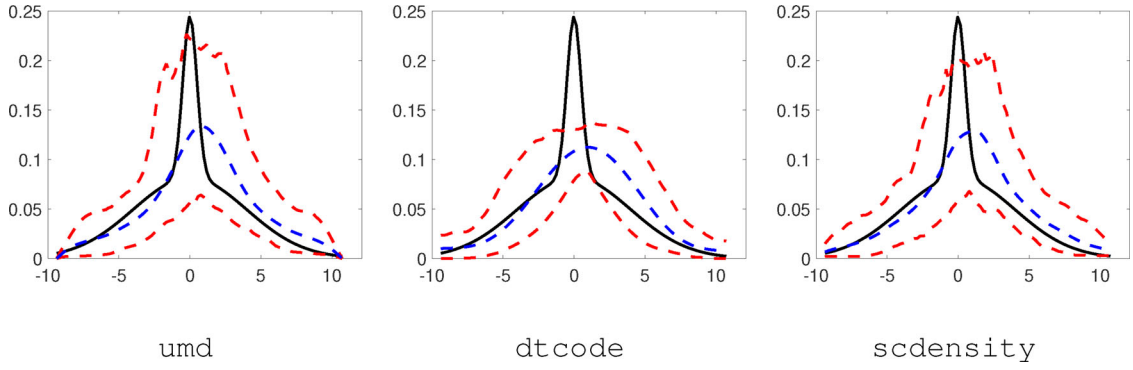

Figure 4. A comparison of the variability of the estimates across different samples for Example 1 from Study 1 at sample size 100. The middle dashed line indicates the average estimate across samples, while the upper and lower dashed lines represent the 95th and 5th quantiles, respectively, of the estimate at the location. The solid line is the true density.

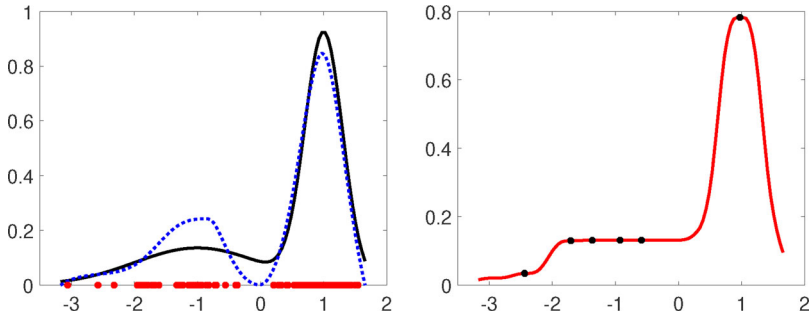

Figure 5. The left panel shows the bimodal example from Study 2, with the true density (solid) and `dtcode` estimate(dotted). The right panel shows the `scdensity` estimate along with local modes. The observations are shown on the *x*-axis.

as a solid line; the estimated density: with best performance as a dashed line; with median performance as a dotted line; and with worst performance as a dashed-dotted line.

In both examples, `dtcode` clearly outperforms `umd` in capturing the sharper features and in stability of performance. On the other hand, the performance of `dtcode` is very similar to the performance of `scdensity` for both the examples. For the kurtotic unimodal example 1, `scdensity` has a slightly better performance, whereas for the contaminated unimodal density estimate 2, `dtcode` is superior. Table 1 in the supplementary materials gives a quantitative analysis.

For the kurtotic unimodal example 1, we also study the pointwise MSE, as shown in Figure 4. The figure shows that across all samples, `dtcode` and the package `scdensity` have a similar overall performance in capturing the location and the height of the mode.

Example 2 is a special case where there are outliers in the data that create the possibility of a small peak near the right boundary. These outliers also affect the boundary estimates, and reflect a spurious mode in the true density. In this example, we see that `dtcode` is very robust to the choice of boundary estimates, replacing the spurious mode with a wide shoulder, as shown in the bottom left panel of Figure 3. For this example, the quantitative performance of `dtcode` is also superior to the other techniques for all sample sizes, as shown in Table 1 of the supplementary materials.

4.2. Study 2

Next, we study a bimodal density: an asymmetric bimodal density given by $p_0 = 1/3\mathcal{N}(-1, 1) + 2/3\mathcal{N}(1, 0.3)$.

We compare the estimation performance of `dtcode` with `scdensity`. Table 2 in the supplementary materials presents a quantitative comparison of different loss functions and the likelihood for this example at different sample sizes. For all sample sizes, the performance of the two approaches is very similar with respect to the loss functions. However, there are some clear advantages to our approach. First, note that the bimodality constraints in Wolters and Braun (2018) are satisfied only on a prespecified finite grid. As a result, the final estimate has spurious modes violating the shape constraint, and thus technically does not belong in the correct shape class; the ability to violate the constraints probably explains the slightly better $\mathbb{L}^2$ errors for its estimates. Second, the estimate itself does not enjoy any statistical optimality. The estimate starts with an unconstrained estimate and obtains the nearest estimate in the correct shape class. For that purpose, it replaces spurious peaks with flat intervals even though the data might suggest otherwise.

Figure 5 illustrates an example of the performance of `dtcode` in comparison with `scdensity`.

The left panel shows the `dtcode` result, while the right panel shows that for `scdensity`. The 100 observations are also shown along the horizontal axis. The quantitative performance of `scdensity` and `dtcode` (shown in Tables 1 and 2 of the supplementary materials) are very similar at all sample sizes. Further investigation reveals that small differences can mostly be attributed to the choice of starting point and the actual optimization algorithm used in our approach, rather than the approach itself. For example, we notice that `scdensity` performs better if we use an external optimization function to obtain the mode locations rather than the approach proposed

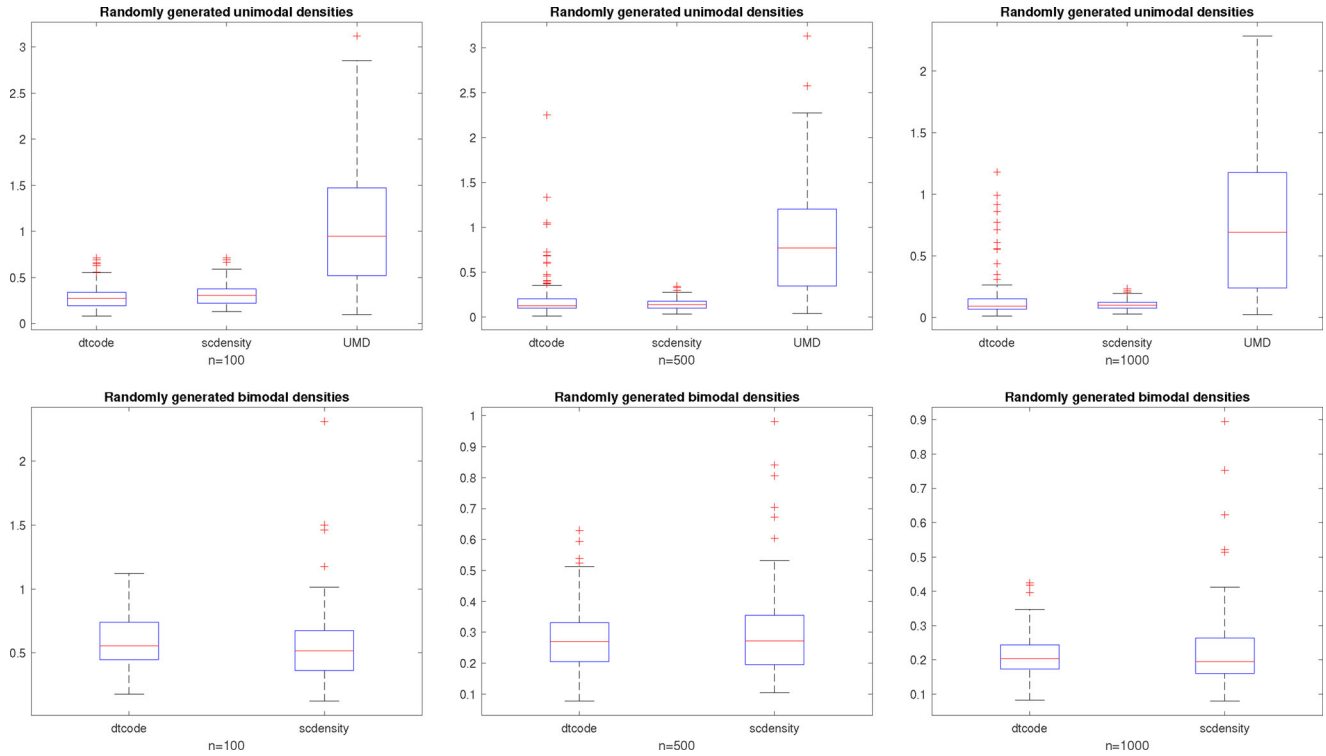

Figure 6. Boxplots of the $\mathbb{L}^2$ norms of the errors of the density estimates found using `dtcode`, `scdensity`, and `umd`, for randomly sampled true densities and different sample sizes. The top row shows the results for Example 1; the bottom row those for Example 2.

in the original article and then used in the `scdensity` package. Also, `dtcode` shows improvement if we choose a more informed starting template shape, such as a kernel density with hand-tuned bandwidth so that the number of modes is correct.

With respect to the shape of the resultant density estimate, however, the `scdensity` estimate does not conform to the available data because the constraint is only satisfied on a prespecified grid. The right panel of Figure 5 shows the `scdensity` estimate along with the local maxima indicated by asterisks. The left modal region is replaced by several small bumps, making it difficult to distinguish a true mode from the constraint violations. We also note that the spurious flat shape in the left tail is probably due to the inbuilt optimization code provided. In comparison, `dtcode` correctly captures the data-sparse region in between the two modal regions and has exactly two modes.

Finally, we emphasize that the shape constraints appear directly in the estimation procedure of Wolters and Braun (2018). This makes the constrained estimation and the nested search for critical points increasingly complex as the modality is increased and makes the approach ill-equipped to handle higher modality constraints. In contrast to `scdensity`, the constraint information in `dtcode` is captured in the initial template function itself, and the subsequent estimation of the transformation is free of the modality information, meaning that the approach scales much better to more general modality constraints.

4.3. Study 3

As an extension of the previous experiments, we now study performance across a range of unimodal and bimodal examples. We do this by averaging performance over random samples from a set of random densities in the appropriate shape family. The true densities themselves are generated randomly as follows:

- Example 1: a unimodal example with random mixing proportions and standard deviations, given by $p_0 = \alpha\mathcal{N}(0,\sigma_1) + (1-\alpha)\mathcal{N}(0,\sigma_2)$, with $\alpha \sim U(0,1)$, $\sigma_1 \sim \max(0.1, \mathcal{N}(0.4, 0.1))$, and $\sigma_2 \sim \max(0.1, \mathcal{N}(3, 0.2))$.
- Example 2: a bimodal example with random mixing proportions, means, and standard deviations, given by $p_0 = \alpha\mathcal{N}(\mu_1,\sigma_1) + (1-\alpha)\mathcal{N}(\mu_2,\sigma_2)$, with $\alpha \sim U(0,1)$, $\sigma_1 \sim \max(0.1, \mathcal{N}(0.75, 0.2))$, $\mu_1 \sim \mathcal{N}(-1, 0.2)$, $\mu_2 \sim \mathcal{N}(0.1, 0.2)$, and $\sigma_2 \sim \max(0.1, \mathcal{N}(0.5, 0.2))$.

Figure 6 shows boxplots of the $\mathbb{L}^2$ norms of the estimation errors for example 1 (top row) and example 2 (bottom row), for three sample sizes. We notice that the `dtcode` estimate is comparable to the `scdensity` estimate at all sample sizes under this measure. Again, as above, our approach is better in terms of the desired shape constraint.

4.4. Study 4

Next, we study pdfs with three and four modes, respectively:

- Example 1: an asymmetric trimodal density with one mode well separated from the other two, given by $p_0 = 1/3\mathcal{N}(-1, 0.25) + 1/3\mathcal{N}(0, 0.25) + 1/3\mathcal{N}(2, 0.3)$.
- Example 2: a four-modal density, given by $p_0 = 0.25\mathcal{N}(-4, 0.5) + 0.25\mathcal{N}(-2, 0.5) + 0.4\mathcal{N}(2, 1) + 0.1\mathcal{N}(5, 0.25)$.

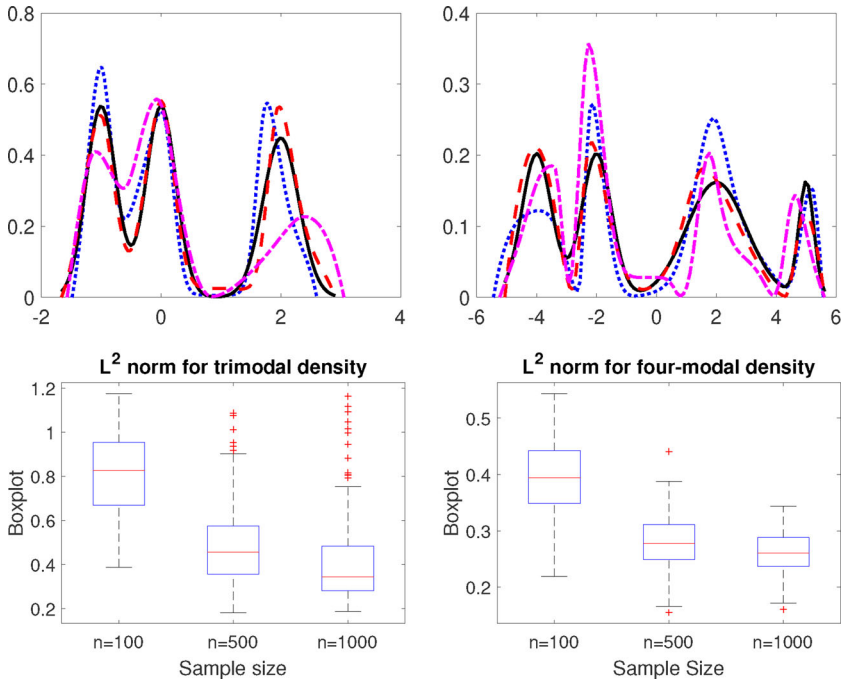

Figure 7. Results of the `dtcode` method on examples 1 (left column) and 2 (right column) of Study 4. The top row shows plots of the true density as a solid line; the estimated density: with best performance as a dashed line; with median performance as a dotted line; and with worst performance as a dashed-dotted line; with performance measure using the $\mathbb{L}^2$ norms of the errors. The bottom row shows boxplots of the $\mathbb{L}^2$ norms of the errors for different samples sizes.

In Figure 7, we use the $\mathbb{L}^2$ norm of the errors calculated for 100 samples of size 100 to study the results obtained with `dtcode` on examples 1 (left column) and 2 (right column). The top row shows plots of the true density as a solid line; the estimated density: with best performance as a dashed line; with median performance as a dotted line; and with worst performance as a dashed-dotted line. The bottom row shows boxplots of the $\mathbb{L}^2$ norms of the errors for different samples sizes. The results show that the performance improves with increasing sample size, in both size and spread of error. Note that we do not compare the `dtcode` results to those of other methods because there is no other method that can handle $M = 3$ or higher.

5. Application to Electricity Consumption Data

Quantification and detection of patterns in electricity consumption curves across households, locations, and seasons, is crucial for planning and forecasting, as discussed in Cordova et al. (2018) and Kwac, Flora, and Rajagopal (2014), among others. Deployment of advanced monitoring systems, including smart meters and synchrophasors, in power distribution networks has created a new paradigm for observing and managing the electric grid, leading to an abundance of consumption data with different levels of granularity. The City of Tallahassee, the capital of Florida, has a Meter Data Management System (MDMS) that stores electricity consumption (kWh) readings from every customer in the city for billing purposes and further analysis. We look at the daily electricity consumption profiles of a randomly chosen de-identified single household in Tallahassee. The dataset was obtained with a Non-Disclosure Agreement with the City of Tallahassee.

The daily consumption patterns show high variability, depending on day of the week, season, and other extraneous factors, even for this single household. We look at the electricity consumption values at different times in a particular day, for four different days, to estimate the daily distribution of electricity consumption. However, one can split the daily consumption profiles into two interpretable clusters: consumption values when the household members are at home versus consumption values when the households are not at home. This suggests that a two-mode constrained density estimation would lend interpretability to the density estimates, which can otherwise be very noisy.

Figure 8 shows the density estimates and the corresponding histograms of electricity consumption for four different days at the randomly chosen household. As expected, in most cases, an unconstrained estimate is too bumpy and uninterpretable. The shape constraint, however, results in much smoother and more interpretable estimates, the two peaks captured by our proposed method aligning well with the major peaks in the histograms.

6. Extension to Conditional Density Estimation

The proposed framework for modality-constrained density estimation extends naturally to modality-constrained *conditional* density estimation. Consider the following setup. Let X be a fixed one-dimensional random variable with a positive density on a fixed interval. Let $Y \sim f_X$, where f_X is an unknown conditional density that changes smoothly with X.

Conditioned on X, Y is assumed to have a univariate, continuous distribution with support on the interval $[A, B]$, with M modes in the interior of $[A, B]$, and $f_X(A) = f_X(B) = 0$. We observe the pairs $\{(Y_i, X_i)\}$, $i = 1, \ldots, n$, and are interested

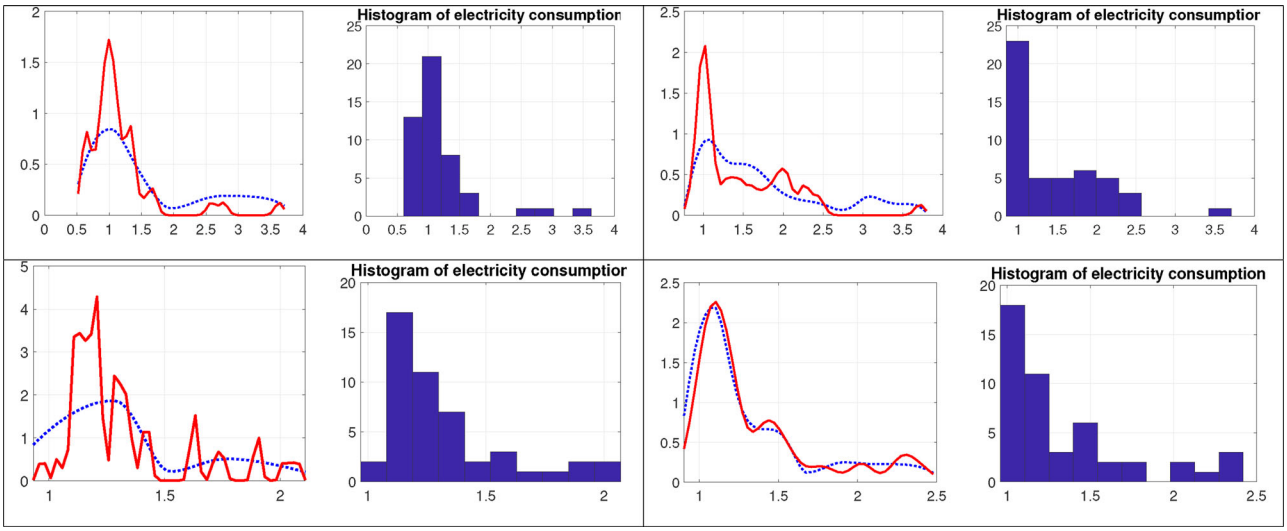

Figure 8. Estimated densities of electricity consumption using the warped approach (dotted) and a kernel (ucv) approach (solid), and the associated histograms of consumption data for four different days.

in recovering the conditional density f_X at a particular location of X, henceforth referred to as x_0. The estimation is again initialized with an M-modal template function $\tilde{p}_\lambda$. However, since f_X varies smoothly with X, we assign more importance to observations closer to the location x_0 than to observations further away, and hence we perform weighted maximum likelihood estimation to find the necessary parameters:

$$(\hat{c}, \hat{\lambda}) = \underset{c \in \mathcal{C}^J, \lambda \in \Lambda_M}{\operatorname{argmax}} \times \left(\sum_{i=1}^{n} \left[\log \left(\tilde{p}_\lambda \left(\gamma_c(x_i) \right) / \int_0^1 \tilde{p}_\lambda \left(\gamma_c(t) \right) \, dt \right) \right] W_{x_0,i} \right), \tag{7}$$

where $W_{x_0,i}$ is the localized weight associated with the ith observation, calculated according to

$$W_{x_0,i} = \frac{\mathcal{N}(||X_i - x_0||_2 / h(x_0); 0, 1)}{\sum_{j=1}^{n} \mathcal{N}(||X_j - x_0||_2 / h(x_0); 0, 1)}. \tag{8}$$

Here $\mathcal{N}(\cdot, 0, 1)$ is the standard normal pdf and $h(x_0)$ is the parameter that controls the relative weights associated with the observations. However, weights defined in this way result in higher bias because information is being borrowed from all observations. To mitigate this, as discussed in an example in Bashtannyk and Hyndman (2001), we allow only a specified fraction of the observations X_i to have a positive weight. Note that 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 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, for the predictors X. A very large value of $h(x_0)$ distributes approximately equal weight over all observations, whereas a very small value considers only the observations in a small neighborhood around x_0. The value of $h(x_0)$ can be chosen via any standard cross-validation-based bandwidth selection method. In our experiments, we use an adaptive bandwidth selection method to save computation time when the predictors are independent of each other. It consists of a 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. (We simply use the `ksdensity` estimate in MATLAB, which chooses the bandwidth optimal for normal densities.) Let h be the fixed bandwidth used.
2. Then, set the bandwidth parameter $h(x_0)$ at location x_0 to be $h(x_0) = h/\sqrt{\hat{K}(x_0)}$.

The intuition behind this choice is that h controls the overall smoothing of the predictor space based on the sample points, while $\sqrt{\hat{K}(x_0)}$ stretches or shrinks the bandwidth at the particular location. In a sparse region, increased borrowing of information from other data points is desirable to reduce the variance of the estimate, whereas in dense regions, reduced borrowing of information from faraway points reduces the bias of the density estimates. A location from a sparse region is expected to have a low density estimate, and a location from a dense region is expected to have a high density estimate. Hence, varying the bandwidth parameter inversely with the density estimate helps adapt to the sparsity around the point of interest. The choice of the adaptive bandwidth parameter is motivated by the variable bandwidth kernel density estimators discussed in Terrell and Scott (1992), Van Kerm (2003), and Abramson (1982), among others. We provide a simulation study in the supplementary materials.

7. Application to Traffic Flow Data

As an application of modality-constrained conditional density estimation, we use the traffic flow data for Californian highways from the package `hdrcde` in R. The scatterplot shown in Figure 9 shows the distinctly bimodal nature of the speed

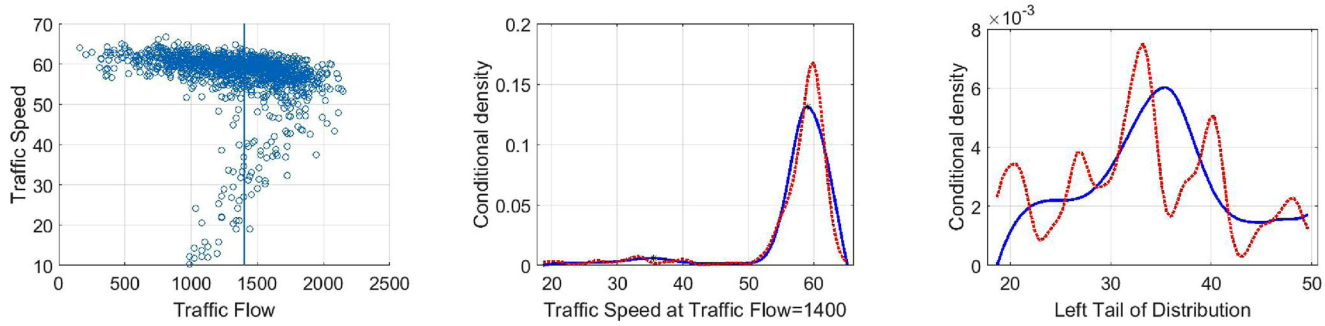

Figure 9. Left: Scatterplot of traffic flow data for Californian highways from the `hdrcde` package in R. Center: Traffic speed density at traffic flow 1400 as estimated by `dtcode` (solid line) and the `NP` package (dotted line). Right: A magnified view of the left part of the center plot.

distribution for traffic flows between 1000 and 1620 vehicles per lane per hour, corresponding to uncongested and congested traffic. This range of traffic flows was studied by Einbeck and Tutz (2006). They noted that beyond a traffic flow of 1620, the regression curves corresponding to uncongested and congested traffic are no longer distinguishable. So, we consider the speed flow in the above range (772 observations), and estimate the density of the speed conditional on a flow of 1400. We use a bimodality constraint on the shape, and our prescribed 50% of the 772 observations. For the tangent space representation, we use up to 6 basis elements.

The middle panel of Figure 9 (solid line) shows the conditional density estimate for flow = 1400 using `dtcode`. The left mode is at 35.56 mph and the right mode is at 59.01 mph. Einbeck and Tutz (2006) obtain a very similar conditional density estimate. The left mode in their case is at 32.65 mph and the right mode is at 59.18 mph. On the other hand, if we find a traditional conditional density estimate using the `NP` package, we find several spurious bumps; this estimate is shown in the middle panel of Figure 9 (dotted line), with a magnified view shown in the right panel. The superfluous bumps are present in the `NP` estimate constructed using 772 observations (not presented), as well as the estimate constructed using only 50% of the observations as in our approach. This results in over-interpreting the tail and consequently a lack of interpretability for the modes themselves. Thus, constraining the number of modes clearly helps with the interpretability of the resulting density.

8. Discussion

Shape-constrained density estimation is a rich problem area that has a broad range of real-world applications, yet has been explored rigorously only in limited cases. Here we have introduced a novel framework, using geometric tools, that enables shape-constrained density estimation using a different notion of shape than studied previously. In our approach, named `dtcode`, a template from the appropriate shape class is deformed using shape-preserving diffeomorphisms of the data domain, the optimal deformation being defined by maximum likelihood. The problem is thereby reframed as one of optimizing over the diffeomorphism group.

The framework is the first in the literature that can perform modality-constrained density estimation for any number of modes. However, from a practical perspective, the performance suffers somewhat when the constrained shape becomes too complex or if the number of modes becomes high (>4). This limitation is due to the current choice of numerical techniques used in optimization over the diffeomorphism group, and because of the choice of basis set used in estimation.

Since this article primarily focuses on the fundamental framework for `dtcode`, it only lightly touches upon or leaves out some associated problems. Examples include the choice of the number of basis elements for the tangent space representation, the choice of the basis itself, estimation of domain boundaries, and the choice of penalty for penalized estimation. These are all interesting problems in their own right, but space limitations force our focus to only the main ideas. Nevertheless, we can make some observations.

- This article uses AIC as the penalty to select the number of basis elements because, in comparison, BIC tends to choose an insufficient number of parameters. However, other model selection techniques can also be investigated.
- Experiments using a Meyer wavelet basis for the tangent space representation yielded results similar to those reported in the article, although the Meyer wavelets seemed to require more observations than the Fourier basis to obtain satisfactory results. Clearly, one can choose different bases and conduct a comparative study of performance. Since the support of the warping functions is compact, we recommend using trigonometric (Fourier and cosine) basis for representation. Please refer to Efromovich (2010) and the references therein for a more detailed discussion on this topic. When the sample size is small, Fourier basis can result in spurious bumps near the boundaries, which is why wavelets may be a good alternative.
- Our article follows Turnbull and Ghosh (2014) in estimating the boundaries, but other choices can be explored as well.
- For conditional density estimation, the weights can be defined using any kernel: the Gaussian kernel (and the $\mathbb{L}^2$-loss function) was only used as an illustration.

An advantage of the proposed framework is that it is easy to extend to conditional density estimation via a weighted maximum likelihood objective function. One potential future direction is to apply this framework to situations where a large number of covariates are present. Currently, the bandwidth parameter is chosen adaptively based on a kernel density estimate at the

location of the (scalar) covariate. The framework can be directly extended to a scenario with d covariates using a d-dimensional kernel density estimate at the location of the predictors. Such an estimate would generically suffer from the curse of dimensionality, but seems valid for applications where only a few of the covariates are relevant. In particular, Wasserman and Lafferty (2006) have developed a technique to shortlist relevant variables and to find corresponding bandwidth parameters. Using these bandwidth parameters, one can redefine the weights and then perform weighted likelihood maximization as before to produce a conditional density estimate.

In conclusion, we have developed a framework for incorporating general modality constraints into a density estimation procedure, while showing very competitive performance on shape constraints already studied in literature. In applications where the data shows modality constraints, the proposed framework will provide accurate and interpretable density estimates that fully respect the constraints in play.

Supplementary Materials

Supplementary materials by section In Section 1 of the supplementary materials, we present a proof of Theorem 1. In Section 2, we discuss the asymptotic properties of our estimator, and present a theorem which provides an upper bound on the convergence rate. We prove this theorem in Section 3. In Section 4, we include tables illustrating the average practical performance for our approach (dtcode), umd, and scdensity, for the examples considered in the simulation study in the main article. In Section 4.1, we discuss the effect of the number of basis elements on the final estimate. In Section 5, we include some examples of general shape-constrained density estimation beyond M-modality, like monotonicity, an upper bound on the number of modes, and so on. In Section 6, we include a simulation study for conditional density estimation. In Section 7, we discuss an application of shape-constrained density estimation to DNA methylation profiles.

Funding

This research was supported in part by grants NSF DMS 1621787 and NSF CCF 1617397 to AS and NSF DMS 1613156 to DP.

References

Abramson, I. S. (1982), "On Bandwidth Variation in Kernel Estimates—A Square Root Law," *The Annals of Statistics*, 10, 1217–1223. [10]

Bashtannyk, D. M., and Hyndman, R. J. (2001), "Bandwidth Selection for Kernel Conditional Density Estimation," *Computational Statistics & Data Analysis*, 36, 279–298. [10]

Bickel, P. J., and Fan, J. (1996), "Some Problems on the Estimation of Unimodal Densities," *Statistica Sinica*, 6, 23–45. [1]

Birge, L. (1997), "Estimation of Unimodal Densities Without Smoothness Assumptions," *The Annals of Statistics*, 25, 970–981. [1]

Brunner, L. J., and Lo, A. Y. (1989), "Bayes Methods for a Symmetric Unimodal Density and Its Mode," *The Annals of Statistics*, 17, 1550–1566. [1]

Cheng, M.-Y., Gasser, T., and Hall, P. (1999), "Nonparametric Density Estimation Under Unimodality and Monotonicity Constraints," *Journal of Computational and Graphical Statistics*, 8, 1–21. [2]

Cordova, J., Sriram, L. M. K., Kocatepe, A., Zhou, Y., Ozguven, E. E., and Arghandeh, R. (2018), "Combined Electricity and Traffic Short-Term Load Forecasting Using Bundled Causality Engine," *IEEE Transactions on Intelligent Transportation Systems*, 20, 3448–3458. [9]

Dasgupta, S., Pati, D., and Srivastava, A. (2021), "A Two-Step Geometric Framework for Density Modeling," *Statistica Sinica* (in press). [2]

Efromovich, S. (2010), "Orthogonal Series Density Estimation," *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 467–476. [5,11]

Einbeck, J., and Tutz, G. (2006), "Modelling Beyond Regression Functions: An Application of Multimodal Regression to Speed-Flow Data," *Journal of the Royal Statistical Society*, Series C, 55, 461–475. [11]

Grenander, U. (1956), "On the Theory of Mortality Measurement: Part II," *Scandinavian Actuarial Journal*, 1956, 125–153. [1]

Hall, P., and Huang, L.-S. (2002), "Unimodal Density Estimation Using Kernel Methods," *Statistica Sinica*, 12, 965–990. [2]

Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R. P., Hong, C., Downey, S. L., Johnson, B. E., Fouse, S. D., Delaney, A., Zhao, Y., and Olshen, A. (2010), "Comparison of Sequencing-Based Methods to Profile DNA Methylation and Identification of Monoallelic Epigenetic Modifications," *Nature Biotechnology*, 28, 1097–1105. [1]

Izenman, A. J. (1991), "Review Papers: Recent Developments in Nonparametric Density Estimation," *Journal of the American Statistical Association*, 86, 205–224. [1]

Kwac, J., Flora, J., and Rajagopal, R. (2014), "Household Energy Consumption Segmentation Using Hourly Data," *IEEE Transactions on Smart Grid*, 5, 420–430. [9]

Lang, S. (2012), *Fundamentals of Differential Geometry* (Vol. 191), New York: Springer. [5]

Meyer, M. C. (2001), "An Alternative Unimodal Density Estimator With a Consistent Estimate of the Mode," *Statistica Sinica*, 11, 1159–1174. [1]

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010), "Detection of Nonneutral Substitution Rates on Mammalian Phylogenies," *Genome Research*, 20, 110–121. [1]

Rao, B. L. S. P. (1969), "Estimation of a Unimodal Density," *Sankhyā: The Indian Journal of Statistics*, Series A (1961–2002), 31, 23–36. [1]

Srivastava, A., and Klassen, E. P. (2016), *Functional and Shape Data Analysis*, New York: Springer. [5]

Terrell, G. R., and Scott, D. W. (1992), "Variable Kernel Density Estimation," *The Annals of Statistics*, 20, 1236–1265. [10]

Turnbull, B. C., and Ghosh, S. K. (2014), "Unimodal Density Estimation Using Bernstein Polynomials," *Computational Statistics & Data Analysis*, 72, 13–29. [1,4,6,11]

Ugray, Z., Lasdon, L., Plummer, J., Glover, F., Kelly, J., and Martí, R. (2007), "Scatter Search and Local NLP Solvers: A Multistart Framework for Global Optimization," *INFORMS Journal on Computing*, 19, 328–340. [6]

Van Kerm, P. (2003), "Adaptive Kernel Density Estimation," *Stata Journal*, 3, 148–156. [10]

Wahba, G. (1981), "Data-Based Optimal Smoothing of Orthogonal Series Density Estimates," *The Annals of Statistics*, 9, 146–156. [4]

Wasserman, L., and Lafferty, J. D. (2006), "Rodeo: Sparse Nonparametric Regression in High Dimensions," in *Advances in Neural Information Processing Systems*, pp. 707–714. [12]

Wegman, E. J. (1970), "Maximum Likelihood Estimation of a Unimodal Density, II," *The Annals of Mathematical Statistics*, 41, 2169–2174. [1]

Wheeler, M., Dunson, D., and Herring, A. (2017), "Bayesian Local Extremum Splines," *Biometrika*, 104, 939–952. [2]

Wolters, M. A., and Braun, W. J. (2018), "A Practical Implementation of Weighted Kernel Density Estimation for Handling Shape Constraints," *Stat*, 7, e202. [2,6,7,8]

Supplementary Materials For Modality-Constrained Density Estimation via Deformable Templates

Sutanoy Dasgupta*

Department of Statistics, Florida State University
and
Debdeep Pati
Department of Statistics, Texas A&M University
and
Ian H. Jermyn
Department of Mathematics and Statistics, Durham University
and
Anuj Srivastava
Department of Statistics, Florida State University

November 22, 2020

*The authors gratefully acknowledge acknowledge NSF DMS 1621787 and NSF CCF 1617397 for supporting this research. Dr. Pati acknowledges NSF DMS 1613156 for supporting this research.

1 Proof of Theorem 1

Theorem 1. *The set of transformations of the set $\mathcal{P}_{M,\lambda}$ by the mapping $\mathcal{P}_{M,\lambda} \times \Gamma \to \mathcal{P}_{M,\lambda}$, given by $(p, \gamma) = \frac{p \circ \gamma}{\int (p \circ \gamma)\, dt}$ is a group action. Furthermore, this action is transitive and free. That is, for any $p, q \in \mathcal{P}_{M,\lambda}$, there exists a unique $\gamma \in \Gamma$ such that $p = (q, \gamma)$.*

Proof. We will call the new function $\tilde{p} \equiv (p, \gamma)$ the *warped* density. To prove the theorem, we first have to establish that the warped density $\tilde{p}$ is indeed in the set $\mathcal{P}_{M,\lambda}$. Note that warping by Γ and the subsequent global scaling do not change the number of modes of p since $\dot{\gamma}$ is strictly positive (by definition). The modes simply get moved to their new locations $\{\tilde{b}_i = \gamma^{-1}(b_i)\}$. Secondly, the height ratio vector of $\tilde{p}$ remains the same as that of p. This is due to the fact that $\tilde{p}(\tilde{b}_i) \propto p(\gamma(\gamma^{-1}(b_i))) = p(b_i)$ and $\tilde{\lambda} = \tilde{p}(\tilde{b}_{i+1})/\tilde{p}(\tilde{b}_1) = p(b_{i+1})/p(b_1) = \lambda$.

Next, we prove the compatibility property that for every $\gamma_1, \gamma_2 \in \Gamma$ and p , we have $(p, \gamma_1 \circ \gamma_2) = ((p, \gamma_1), \gamma_2)$. This property holds because

$$((p, \gamma_1), \gamma_2) = \frac{\frac{p \circ \gamma_1}{\int (p \circ \gamma_1)\, ds} \circ \gamma_2}{\int \left(\frac{p \circ \gamma_1}{\int (p \circ \gamma_1)\, ds} \circ \gamma_2 \right) dt} = \frac{p \circ \gamma_1 \circ \gamma_2}{\int (p \circ \gamma_1 \circ \gamma_2)\, dt} = (p, \gamma_1 \circ \gamma_2)\ .$$

Finally, we prove that the action is transitive and free: given $p, \tilde{p} \in \mathcal{P}_{M,\lambda}$, there exists a unique $\gamma_0 \in \Gamma$ such that $\tilde{p} = (p, \gamma_0)$. Let h_p be the height of the first mode of p and let $h_{\tilde{p}}$ be the height of the first mode of $\tilde{p}$. Then, define two nonnegative functions according to $g = p/h_p$ and $\tilde{g} = \tilde{p}/h_{\tilde{p}}$. Note that the height of both their first modes is 1 and the height vector for the interior critical points is still λ. Also, let the critical points of p and $\tilde{p}$ (and hence g and $\tilde{g}$, respectively) be located at b_i and $\tilde{b}_i$ respectively, for $i = 0, \cdots, 2M$. Since the modes are well defined, the function g is piecewise strictly monotonic 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 unique (and continuous) inverse of

g, termed g_t^{-1}. Then, $\gamma_0(x) = g_t^{-1}\big(\tilde{g}(x)\big), x \in [\tilde{b}_t, \tilde{b}_{t+1}]$ is such that $(g \circ \gamma_0) = \tilde{g}$ and hence $(p, \gamma_0) = \tilde{p}$.

□

2 Asymptotic Convergence Results

In this section, we derive the asymptotic rate of convergence to the true underlying density p_0 of the (maximum likelihood) density estimate $\hat{p}$ described by equation 5 in Section 3.2 of the main paper. To do this, we use the theory of sieve maximum likelihood estimation as in Wong and Shen [1995]. Let $\mathcal{P}$ denote the set of M-modal continuous densities on $[0, 1]$ strictly positive in $(0, 1)$ and zero at the boundaries (we drop the previously-used subscript M for simplicity).

- Assumption 1: $p_0 : [0, 1] \to \mathbb{R}_+$ is continuous, strictly positive on $(0, 1)$, and $p_0(0) = p_0(1) = 0$.
- Assumption 2: p_0 has M modes which lie in $(0, 1)$.
- Assumption 3: p_0 either belongs to Hölder or Sobolev space of order β.

Let n be the number of available observations. Let $\mathcal{P}_n$ be the approximating space of $\mathcal{P}$ when using $J = k_n$ basis elements for the tangent space $T_{\mathbf{1}}(\mathbb{S}^+_\infty)$, where k_n is some function of the number of observations n. Let η_n be a sequence of positive numbers converging to 0. Let $Z_i \in (0, 1)$ be the n observed data points. We call an estimator $\hat{p} : [0, 1] \to \mathbb{R}_+$ an η_n sieve MLE if

$$\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}(Z_i) \geq \sup_{p \in \mathcal{P}_n} \frac{1}{n}\sum_{i=1}^{n} \log p(Z_i) - \eta_n$$

In the proposed method, $\hat{p}$ is defined such that $\frac{1}{n}\sum_{i=1}^{n}\log\hat{p}(Z_i)$ is exactly $\sup_{p\in\mathcal{P}_n}\frac{1}{n}\sum_{i=1}^{n}\log p(Z_i)$. Therefore, $\hat{p}$ is a sieve MLE with $\eta_n \equiv 0$. Let $\|\cdot\|_r$ denote $\mathbb{L}^r$ norm between functions. The following theorem states the asymptotic convergence rate for the sieve MLE $\hat{p}$.

Theorem 2. *Let $\epsilon_n^* = M_1 n^{-\beta/(2\beta+1)}\sqrt{\log n}$ for some constant M_1. If p_0 satisfies Assumptions 1, 2 and 3; and $\hat{p}$ is the sieve MLE described according to* (**??**) *in Section* **??**, *then there exist constants C_1 and C_2 such that*

$$P(\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \geq \epsilon_n^*) \leq 5\exp\left\{-C_2 n(\epsilon_n^*)^2\right\} + \exp\left\{-\frac{1}{4}nC_1(\epsilon_n^*)^2\right\}. \tag{1}$$

The proof hinges on establishing an equivalence between the density space $\mathcal{P}$ and the parameter space. That is, we show that if the estimated parameter is "close" to the parameter corresponding to the true density, then the corresponding estimated density is also "close" to the true density. The statement is formally stated and proved in the next section.

3 Proof of Theorem 2

3.1 Preliminaries and Auxiliary results

First we set some notation and some preliminary definitions. M is always used to represent the number of modes. Let n be the sample size. Let g_λ denote the M-modal template defined earlier as a function of λ, except that we now parameterize the boundary values of g_λ by a positive number ω, and henceforth denote the template as g_λ^ω. Here λ denotes the vector $(\lambda_1 \cdots \lambda_{2M-2})$, corresponding to the $2M-2$ height ratios of the last $2M-2$ critical points with respect to the

first critical point. Let k_n be the number of basis elements used for approximating the warping function γ. Let $c = (c_1, \cdots, c_{k_n})$ be a corresponding coefficient vector. Now, define $\theta_n = (c_1, \cdots c_n, \lambda_1, \cdots, \lambda_{2M-2})$. In what follows, c is used to represent coefficient vectors. B_i denotes the i^{th} basis element for the tangent space representation of warping functions. γ_c is used to represent the warping function corresponding to the coefficient vector c. $l_1, l_2, \cdots, C, C_1, \cdots$ represent specific constants. $M_0, M_1, M_2, \cdots$ represent generic constants that can change values from step to step but are otherwise independent of other terms.

Let $\lambda_0 \in \mathbb{R}^{2M-2}$ be the *height ratio vector* for p_0, as defined in Section 2. Then from Theorem 1 there exists an infinite dimensional c_0 such that p_0 can be represented as

$$p_0 = (g^0_{\lambda_0} \circ \gamma_{c_0}) / \int_0^1 (g^0_{\lambda_0} \circ \gamma_{c_0}) dt.$$

Note that for each $t \in [0,1]$, $\| \sum_{i=1}^{\infty} c_i B_i(t) \| = \sqrt{\int \left(\sum_{i=1}^{\infty} c_i B_i(t) \right)^2} < 2\pi$. This corresponds to $\max_{1 \le j \le k_n} |c_{0j}| < l_0$ and thus $|c_{0i}| < l_0$ for all i, for some l_0. Then the parameter space for $\mathcal{P}$ is $\Theta = \{(c, \lambda) : c \in [-l_0, l_0]^{\infty}, \lambda \in \Lambda \subset (0, \infty)^{2M-2}\}$. Note that $\omega = \omega(n) = \Omega / \log n$.where Ω is a constant. Let $r_n^u = \Omega_1 \log n$ and $r_n^l = \Omega_1 / \log n$ where $\Omega_1 < \Omega$ is some constant. Define $\mathcal{P}_n$ as the approximating space of densities for $\mathcal{P}$. Define $\Theta_n = \{\theta_n = (c, \lambda) : c \in [-l_0, l_0]^{k_n}, \lambda \in (r_n^l, r_n^u]^{2M-2}\}$ as the parameter space for the approximating space $\mathcal{P}_n$. Then $\mathcal{P}_n = (g^{\omega}_{\lambda} \circ \gamma_c) / \int_0^1 (g^{\omega}_{\lambda} \circ \gamma_c) dt$ where $\theta_n = (c, \lambda) \in \Theta_n$. We use the method of sieve maximum likelihood estimation to obtain the estimate in the approximating space $\mathcal{P}_n$ of $\mathcal{P}$ and to derive an upper bound of the convergence rate of the density estimate to the final density.

We call a finite set $\{(f_j^L, f_j^U), j = 1, \ldots, N\}$ a Hellinger u-bracketing of $\mathcal{P}_n$ if $\| {f_j^L}^{1/2} - {f_j^U}^{1/2} \|_2 \le u$ for $j = 1, \ldots, N$, and for any $p \in \mathcal{P}_n$, there is a j such that $f_j^L \le p \le f_j^U$. Let $H(u, \mathcal{P}_n)$ denote the Hellinger metric entropy of $\mathcal{P}_n$, defined as the logarithm of the cardinality of the u-bracketing

of $\mathcal{P}_n$ of the smallest size. To control the approximation error of $\mathcal{P}_n$ to $\mathcal{P}$, Wong and Shen [1995] introduced a family of discrepancies $\delta_n(p_0, \mathcal{P}_n) = \inf_{p\in\mathcal{P}_n}\rho(p_0, p)$, called the ρ-approximation error at p_0. Controlling $\delta_n(p_0, \mathcal{P}_n)$ is necessary for obtaining results on the convergence rate for sieve MLEs. We follow Wong and Shen [1995] to introduce a family of indexes of discrepency in order to formulate the condition on the approximation error of $\mathcal{P}_n$. Let

$$Z_\alpha(x) = \begin{cases} (1/\alpha)[x^\alpha - 1], -1 < \alpha < 0 \text{ or } 0 < \alpha \leq 1 \\ \log x, \text{ if } \alpha = 0 + . \end{cases}$$

Set $x = p_0/p$ and define $\rho_\alpha(p_0, p) = E_p Z_\alpha(X) = \int p_0 Z_\alpha(p_0/p)$. We define $\delta_n(\alpha) = \inf_{p\in\mathcal{P}_n}\rho_\alpha(p_0, p)$. For our purposes we set $\alpha = 1$. Thus we have $\delta_n(1) = \inf_{p\in\mathcal{P}_n}\int (p_0 - p)^2/p$.

Let f_1 and f_2 be two densities in $\mathcal{P}_n$. Let $\theta_1 = (c_1, \lambda_1)$ and $\theta_2 = (c_2, \lambda_2)$ be the corresponding parameters. g_1^ω and g_2^ω be the corresponding templates. Let M be the number of modes and γ_1 and γ_2 be the warping functions corresponding to the coefficients. Then we have

Lemma 1. $|f_1 - f_2| \leq M_0 \sum_{i=1}^{k_n+2M-2} |\theta_{1i} - \theta_{2i}|$, *for some constant* $M_0 > 0$.

Proof. First, following the steps of Dasgupta et al. [In press] we observe that $|\gamma_1(t) - \gamma_2(t)| < M_2 \sum_{i=1}^{k_n} |c_{1i} - c_{2i}| < M_1 \sum_{i=1}^{k_n+2M-2} |\theta_{1i} - \theta_{2i}|$ since the c_i's are simply the first few coordinates of θ. Next, observe that $|g_1^\omega \circ \gamma_1 - g_2^\omega \circ \gamma_2| \leq |g_1^\omega \circ \gamma_1 - g_1^\omega \circ \gamma_2| + |g_1^\omega \circ \gamma_2 - g_2^\omega \circ \gamma_2|$. By construction, g_1^ω is Lipschitz continuous, and hence $|g_1^\omega \circ \gamma_1 - g_1^\omega \circ \gamma_2| \leq M_2|\gamma_1 - \gamma_2| \leq M_3 \sum_{i=1}^{k_n+2M-2} |\theta_{1i} - \theta_{2i}|$. Now, we have $|g_1^\omega \circ \gamma_2 - g_2^\omega \circ \gamma_2| \leq \max_{1\leq i\leq(2M-2)} |\lambda_{1i} - \lambda_{2i}| \leq M_2 \sum_{i=1}^{k_n+2M-2} |\theta_{1i} - \theta_{2i}|$. Thus, it follows that $|g_1^\omega \circ \gamma_1 - g_2^\omega \circ \gamma_2| \leq M_1 \sum_{i=1}^{k_n+2M-2} |\theta_{1i} - \theta_{2i}|$. Using the above observations, we prove the Lemma.

Let $I_1 = \int_0^1 g_1^\omega \circ \gamma_1 dt$ and $I_2 = \int_0^1 g_2^\omega \circ \gamma_2 dt$. Then we have $0 < r_n^l = \min(\inf_i \lambda_{ki}, g_1^\omega(0), g_1^\omega(1)) <$

$I_k < \max(1, \sup_i \lambda_{ki}) = r_n^u$ for $k = 1, 2$. Now, we have

$$|f_1 - f_2| = \left| \frac{(g_1^\omega \circ \gamma_1) I_1 - (g_2^\omega \circ \gamma_2) I_2}{I_1 I_2} \right| = \left| \frac{(g_1^\omega \circ \gamma_1) I_1 - (g_2^\omega \circ \gamma_2) I_1}{I_1 I_2} + \frac{(g_2^\omega \circ \lambda_2)(I_1 - I_2)}{I_1 I_2} \right|.$$

Hence,

$$|f_1 - f_2| \leq \left| \frac{(g_1^\omega \circ \gamma_1) - (g_2^\omega \circ \gamma_2)}{I_2} \right| + \frac{(g_2^\omega \circ \lambda_2)}{I_1 I_2} |I_1 - I_2| \leq M_1 \sum_{i=1}^{k_n + 2M - 2} |\theta_{1i} - \theta_{2i}| + \frac{(g_2^\omega \circ \lambda_2)}{I_1 I_2} |I_1 - I_2|$$

where the last inequality is obtained using the fact that I_2 is a finite positive number. Now, $(g_2^\omega \circ \lambda_2) < \max(1, r_n^u)$. Thus $(g_2^\omega \circ \lambda_2)/I_1 I_2$ is bounded above by $r_n^{-2l} \max(1, r_n^u)$. Next, it is easy to check that $|I_1 - I_2| \leq M_1 \|(g_1^\omega \circ \gamma_1) - (g_2^\omega \circ \gamma_2)\|_\infty \leq M_2 \|(g_1^\omega \circ \gamma_1) - (g_2^\omega \circ \gamma_2)\|_1$. Thus we have $|f_1 - f_2| \leq M_0 \sum_{i=1}^{k_n + 2M - 2} |\theta_{1i} - \theta_{2i}|$. □

Remark 1. *It follows that* $H(f_1, f_2) < l_1 \sqrt{\|f_1 - f_2\|_1} < l_1 \sqrt{\sum_{i=1}^{k_n + 2M - 2} |\theta_{1i} - \theta_{2i}|} < l_1 \sqrt{\max_{1 \leq j \leq k_n + 2M - 2} |\theta_{1j} - \theta_{2j}|}$ *for some fixed* $l_1 > 0$ *where* $H(f_1, f_2)$ *is the Hellinger metric between two densities* f_1 *and* f_2.

Corollary 1. *Let* p_0 *be the true density. If* $k_n \sim n^{1/(2\beta+1)}$, *then asymptotically* $\inf_{f \in \mathcal{P}_n} \|p_0 - f\|_\infty \sim n^{-\beta/(2\beta+1)}$ *where* β *is the order of the Sobolev space.*

This corollary follows from standard approximation results in $\mathbb{L}^2$ basis (e.g. Fourier) of Hölder functions of order β. For a detailed discussion please refer to Triebel [2006].

Lemma 2. *There exists positive constants* C_3, C_4, *such that for some positive* $\epsilon < 1$,

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} H^{1/2}(\frac{u}{C_3}, \mathcal{P}_n) du \leq C_4 n^{1/2} \epsilon^2 \tag{2}$$

Proof. The u/C_3-cover of a set T with respect to a metric ρ is a set $\{f^1, \ldots, f^N\} \subset T$ such that for each $f \in T$, there exists some $i \in \{1, \ldots, N\}$ with $\rho(f, f^i) \leq u/C_3$. The covering number N is the cardinality of the smallest delta cover. Then $\log(N)$ is the metric entropy

for T. First we bound the metric entropy for $\mathcal{P}_n$. Let us consider a fixed $f_1, f_2 \in \mathcal{P}_n$. We choose the Hellinger metric for the space $\mathcal{P}_n$ so that we can borrow results directly from Wong and Shen [1995]. We note that $H(f_1, f_2) \leq l_1\sqrt{\max_{1\leq j\leq k_n+2M-2}|\theta_{1j}-\theta_{2j}|}$ for some $l_1 > 0$ following the Remark 1. So finding a u/C_3 covering for $\mathcal{P}_n$ using Hellinger metric is equivalent to finding an $l_1\sqrt{u/C_3}$ covering for the space of parameters $\Theta_n = \{\theta_n = (c, \lambda) : c \in [-l_0, l_0]^{k_n}, \lambda \in (r_n^l, r_n^u]^{2M-2}\}$ using L_∞ norm for euclidean vectors. The $l_1\sqrt{u/C_3}$ covering number for Θ_n using L_∞ norm is $(\frac{2l_0}{l_1}\sqrt{C_3/u})^{k_n}(\frac{(r_n^u-r_n^l)}{l_1}\sqrt{C_3/u})^{(2M-2)}$. This is obtained by partitioning the intervals $[-l_0, l_0]$ and $[r_n^l, r_n^u]$ into pieces of length $l_1\sqrt{u/C_3}$ corresponding to individual coordinates and thus obtaining the partition of Θ_n through cross product. Then in each equivalent class of the partition of Θ_n we have $\|\theta_1 - \theta_2\|_\infty \leq l_1\sqrt{u/C_3}$. Thus the covering number is $(\frac{2l_0}{l_1}\sqrt{C_3/u})^{k_n}(\frac{(r_n^u-r_n^l)}{l_1}\sqrt{C_3/u})^{(2M-2)} < (\frac{2l_0}{l_1}\sqrt{C_3/u})^{k_n}(\frac{r_n^u}{l_1}\sqrt{C_3/u})^{(2M-2)} < (\frac{2l_0\sqrt{C_3}+r_n^u\sqrt{C_3}}{l_1\sqrt{u}})^{(k_n+2M-2)} = N$, say. So the metric entropy for $\mathcal{P}_n$, $H(u/C_3, \mathcal{P}_n)$ is bounded by $\log(N) = (k_n + 2M - 2)\log(\frac{2l_0\sqrt{C}_3+r_n^u\sqrt{C}_3}{l_1\sqrt{u}})$.

Now, note that $r_n^u = \Omega_1 \log n$. Then there exists a constant l_2 such that $2l_0\sqrt{C}_3 + r_n^u\sqrt{C}_3 < l_2 r_n^u$. Also, let $k_n = n^{1/(2\beta+1)} = n^\Delta$. Then there exists a constant l_3 such that $k_n + 2M - 2 < l_3 k_n$.Thus we have, $\log(N) < l_3 k_n \log(\frac{r_n^u l_2}{l_1\sqrt{u}})$. Thus we have $H^{1/2}(u/C_3, \mathcal{P}_n) < \sqrt{\log N} < \sqrt{l_3 k_n \log(\frac{r_n^u l_2}{l_1\sqrt{u}})}$. Let $l_4 = 2^8 l_2/l_1$. Hence,

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} H^{1/2}(u/C_3, \mathcal{P}_n) < \sqrt{l_3 n^\Delta}\int\sqrt{\log\frac{l_2 r_n^u}{l_1\sqrt{u}}} < \sqrt{l_3 n^\Delta \log\frac{l_4 r_n^u}{\epsilon^2}}(\sqrt{2}\epsilon - \frac{\epsilon^2}{2^8}) < \sqrt{2l_3\epsilon^2 n^\Delta \log\frac{l_4 r_n^u}{\epsilon^2}}$$

Then as $\epsilon \uparrow 1$, there exists a constant C_4 such that $\sqrt{2l_3\epsilon^2 n^\Delta \log\frac{l_4 r_n^u}{\epsilon^2}} \leq C_4 n^{1/2}\epsilon^2$. Thus there exists an $\epsilon < 1$ for which (2) holds. □

Now we are ready to provide the proof of Theorem 2.

3.2 Main Proof

Theorem 1 of Wong and Shen [1995] states that, if (2) holds for some $\epsilon < 1$, then there exists constants C_1, C_2 such that the following likelihood surface inequality holds.

$$P^*\left(\sup_{\{\|p^{1/2}-p_0^{1/2}\|_2 \geq \epsilon, p\in\mathcal{P}_n\}} \prod_{i=1}^{n} p(Y_i)/p_0(Y_i) \geq \exp(-C_1 n\epsilon^2)\right) \leq 4\exp(-C_2 n\epsilon^2) \tag{3}$$

Next we derive an expression for an upper bound of the smallest $\epsilon < 1$ that satisfies (2). Let the smallest ϵ, denoted by ϵ_n be of the form $\sqrt{l_4} n^{-\eta}(\log n)^{\nu}$. Then $\log \frac{l_4 r_n^u}{{\epsilon_n}^2} = \log n^{2\eta}(\log n)^{1-2\nu} = (2\eta)\log n + (1-2\nu)\log\log n < (\delta + 2\eta)\log n$. Thus an upper bound for ϵ_n can be obtained by solving

$$\sqrt{2l_3 l_4 n^{-2\eta}(\log n)^{2\nu} n^{\Delta}(2\eta\log n + (1-2\nu)\log\log n)} = C_4 n^{1/2} l_4 n^{-2\eta}(\log n)^{2\nu}.$$

Setting $\nu = 1/2$, and noting that $\Delta = 1/(2\beta+1)$ we get $\eta = \beta/(2\beta+1)$. Thus, $\epsilon_n = \sqrt{l_4} n^{\frac{-\beta}{2\beta+1}}\sqrt{\log n}$ is an upper bound of the smallest ϵ that satisfies (2).

Consider the family of discrepancies $\delta_n(\alpha)$ with $\alpha = 1$. Let the true density be p_0 with corresponding parameters c_0 and λ_0. $\delta_n(1) = \inf_{p\in\mathcal{P}_n} \rho_1(p_0, p). = \inf_{p\in\mathcal{P}_n} \int (p_0 - p)^2/p$. Let $p_1 = \operatorname{arginf}_{p\in\mathcal{P}_n} \int (p_0 - p)^2/p$. Then $\delta_n(1) < \|p_0 - p_1\|_\infty^2 \int 1/f < \|p_0 - p_1\|_\infty^2 \min(r_n^l, \omega) \sim n^{-2\beta/(2\beta+1)}\log n$.

Let C_1, C_2 satisfy (3). Define as in Theorem 4 of Wong and Shen [1995],

$$\epsilon_n^* = \begin{cases} \epsilon_n, \text{ if } \delta_n(1) < \frac{1}{4} C_1 {\epsilon_n}^2, \\ (4\delta_n(1)/C_1)^{1/2}, \text{ otherwise.} \end{cases}$$

Note that $\delta(1)$ and ϵ_n are equal up to constants. It follows from Theorem 4 of Wong and Shen [1995], that

$$P(\|\hat{p}^{1/2} - p_0^{1/2}\|_2 \geq \epsilon_n^*) \leq 5\exp\left\{-C_2 n(\epsilon_n^*)^2\right\} + \exp\left\{-\frac{1}{4} n C_1 (\epsilon_n^*)^2\right\}.$$

4 Tables for comparing performance with other methods

Table 1: A quantitative comparison of the `dtcode`, `umd`, and `scdensity` results for unimodal datasets.

Example:		Symmetric Unimodal						Contaminated Unimodal					
Method:		`dtcode`		`umd`		scdensity		`dtcode`		`umd`		scdensity	
n	Norm	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
100	$\mathbb{L}^1$	1.19	0.30	1.58	0.22	**1.10**	0.26	**3.06**	1.56	6.66	1.44	3.14	1.09
	$\mathbb{L}^2$	0.19	0.06	0.28	0.01	**0.18**	0.05	**0.44**	0.23	0.95	0.24	0.49	0.17
	$\mathbb{L}^\infty$	0.08	0.03	0.12	0.01	**0.07**	0.03	**0.11**	0.06	0.15	0.05	0.14	0.06
500	$\mathbb{L}^1$	0.57	0.11	1.19	0.11	**0.56**	0.10	**1.23**	0.52	3.42	0.87	1.55	0.45
	$\mathbb{L}^2$	0.1	0.02	0.23	0.01	**0.09**	0.02	**0.19**	0.09	0.51	0.16	0.26	0.08
	$\mathbb{L}^\infty$	**0.04**	0.01	0.11	0.01	**0.04**	0.01	**0.05**	0.03	0.15	0.05	0.09	0.03
1000	$\mathbb{L}^1$	0.48	0.29	1.13	0.06	**0.40**	0.07	**0.83**	0.32	3.15	0.89	1.06	0.29
	$\mathbb{L}^2$	0.08	0.06	0.22	0.01	**0.06**	0.01	**0.12**	0.06	0.46	0.09	0.18	0.06
	$\mathbb{L}^\infty$	0.04	0.34	0.11	0.01	**0.03**	0.01	**0.04**	0.13	0.05	0.05	0.06	0.03

Table 2: A quantitative comparison of the performances of `dtcode` and `scdensity` for the simulated bimodal example.

Method:		`dtcode`		`scdensity`	
n	Norm	Mean	Std Dev	Mean	Std Dev
100	$\mathbb{L}^1$	4.5947	1.0687	**4.2034**	1.2635
	$\mathbb{L}^2$	0.6438	0.1692	**0.6285**	0.2050
	$\mathbb{L}^\infty$	**0.1870**	0.0725	0.2122	0.0850
	LogLike	**-105.8070**	11.0621	-106.7010	10.8278
500	$\mathbb{L}^1$	**2.1147**	0.5589	2.1785	0.4888
	$\mathbb{L}^2$	**0.3225**	0.0981	0.3371	0.0880
	$\mathbb{L}^\infty$	**0.1058**	0.0450	0.1297	0.0417
	LogLike	**-554.5179**	21.3556	-555.3835	21.2605
1000	$\mathbb{L}^1$	1.6946	0.4362	**1.6410**	0.3210
	$\mathbb{L}^2$	0.2694	0.0831	**0.2489**	0.0628
	$\mathbb{L}^\infty$	0.1005	.0437	**0.0947**	0.0339
	LogLike	**-1108.3**	33.8369	-1108.5	32.4121

4.1 Effect of number of basis elements

Figure 1 shows some effects of increasing the number of Fourier basis elements with a bimodal example (sample size 100). The optimal number of basis elements here is four (dashed line). If we use eight elements (dotted line) or 14 elements (dashed-dotted line) instead, the estimator tends to place sharp bumps at isolated points. For 14 elements, the overall estimate also suffers as

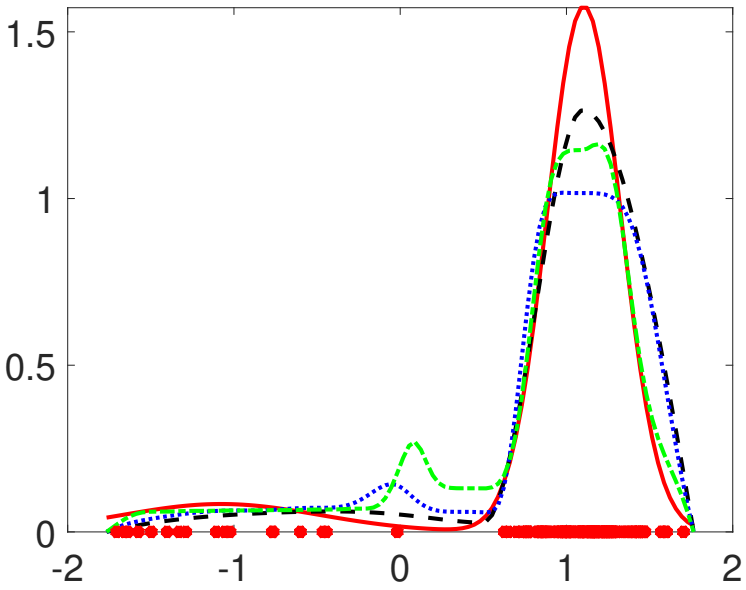

Figure 1: An example showing the effects of increasing the number of basis elements. The true density is shown as a solid line; the estimate with 4 elements as a dashed line; with 8 elements as a dotted line; with 14 elements as a dashed-dotted line.

we see a flatter shape at the mode. This motivates the need for a penalized likelihood to estimate the number of modes.

5 Extension to general constraints

Up to this point we have restricted ourselves to density estimates which are zero at the boundary, even though the true density might not be exactly zero. Also the estimation has inherently assumed that the M modes lie in the interior of the support and not on the boundary. As indicated in the simulation studies, the method has very good numerical performance for densities which decay at the boundaries. However, the proposed framework allows a easy extension to more general constraints: these are discussed in the following sections.

5.1 Densities with non-zero boundary values

The framework can be extended to estimate densities which may have (1) modes located at the boundaries, or (2) compact support with non-zero values at the boundaries, by simply considering the height ratios at the boundaries as extra parameters. The rest of the procedure remains the same. Another special example are monotone densities, where the mode is at one of the boundaries. In such a scenario, one can construct the template by setting the modal value of g to be 1, and then estimate the other boundary value λ_1 with the appropriate constraint.

Further, suppose a density has a flat spot at a modal (or antimodal) location. This indicates that the mode is not well defined but is actually an interval. The framework, in principle, accommodates such information by simply adding a flat spot in the template function at the desired location. Thus, we can extend the idea of the 'shape' of a continuous density function to an ordered sequence of increasing, decreasing, or flat pieces that form the entire density function. For example, a simple bimodal density function can be identified with the sequence *increasing-decreasing-increasing-decreasing*. A function with a unique modal interval can be described as *increasing-flat-decreasing*. If this sequence is known, then simply constructing a template with the same sequence allows us to provide a maximum likelihood density estimate within the class of densities corresponding to that shape sequence.

We consider three examples:

1. a monotonically decreasing density function, given by $p_0 \propto \mathcal{N}(0, 0.4) I_{x\in[0,1]}$, and zero otherwise;

2. a density function with a flat modal region, given by $p_0 \propto x I_{x\in[0,1/3]} + 1/3 I_{x\in[1/3,2/3]} + (1-x) I_{x\in[2/3,1]}$, and zero otherwise;

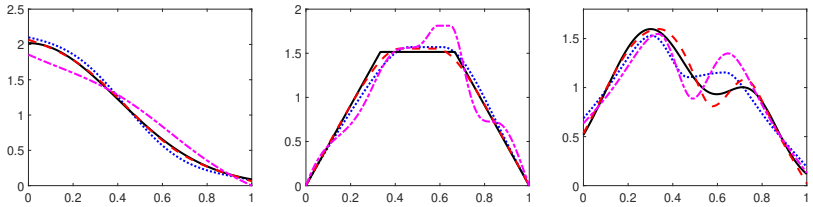

Figure 2: The left panel shows the true (solid line), best (dashed line), median (dotted line), and worst (dashed-dotted line) performances out of 100 samples (according to $\mathbb{L}^2$ norm) of size 500 from the monotonically decreasing density; the middle panel shows the same for sample size 500 from the density with a flat region; the right panel shows the same for sample size 1000 from the truncated bimodal density.

3. a bimodal density function truncated to $[0, 1]$, given by $p_0 \propto 3/4\mathcal{N}(0.3, 0.2^2)I_{[0,1]} + 1/4\mathcal{N}(0.75, 1/8^2)I_{[0,1]}$.

As before, we use the MATLAB function `fmincon` for optimization. Figure 2 shows the results. The left panel shows the true (solid line), best (dashed line), median (dotted line), and worst (dashed-dotted line) performances out of 100 samples (according to $\mathbb{L}^2$ norm) of size 500 from the monotonically decreasing density; the middle panel shows the same for sample size 500 from the density with a flat region; the right panel shows the same for sample size 1000 from the truncated bimodal density.

5.2 Upper bound on modes

In many situations, the exact number of modes might be unclear, but one can put an upper bound on the number of modes in the true density. The proposed framework extends naturally to constraints giving an upper bound $\tilde{M}$ on the number of modes, producing an estimate which has

$m \leq \tilde{M}$ modes, as follows. Note that if the template has M modes, then subsequent composition with a diffeomorphism and renormalization (the group action) cannot create new modes. However, the number M of modes is maintained by the inequalities imposed on the height ratio vector, which do not allow the height ratio of a mode to become less than the adjacent antimodes, and vice versa. If these inequality constraints on the height ratio vectors are relaxed, then one can obtain any $m \leq M$ modes. However, a detailed analysis of this topic is beyond the scope of this paper.

6 Simulation Study For Conditional Density Estimation

We consider two illustrative examples:

1. a unimodal conditional density, given by $X \sim \mathcal{N}(0, 10)$ and $Y|X \sim \text{DExp}((2X-1)^2, 1)$;
2. a bimodal conditional density, given by $X \sim \mathcal{N}(0, 1)$, and $Y|X \sim 0.5\,\mathcal{N}(X - 1.5, 0.5^2) + 0.5\,\mathcal{N}(X + 1.5, 0.5^2)$.

In both cases, we study 100 samples of size 100 and 1000, and compute the conditional density at the 25^{th}, 50^{th}, and 75^{th} quantile of the predictor support. Figure 3 and Figure 4 illustrate the true (solid), best (dashed), worst (dashed-dotted), and median (dotted) performances among the 100 samples for the unimodal example and the bimodal example respectively.

For sample size 100, the performance is slightly unstable and the worst performance often has a bias and is wiggly in nature. Naturally, for larger sample size 1000, the results are much more stable. Also noteworthy is the more pronounced bias for the conditional densities evaluated at the 25^{th} and 75^{th} quantiles, because of the borrowing of information via weighted likelihood

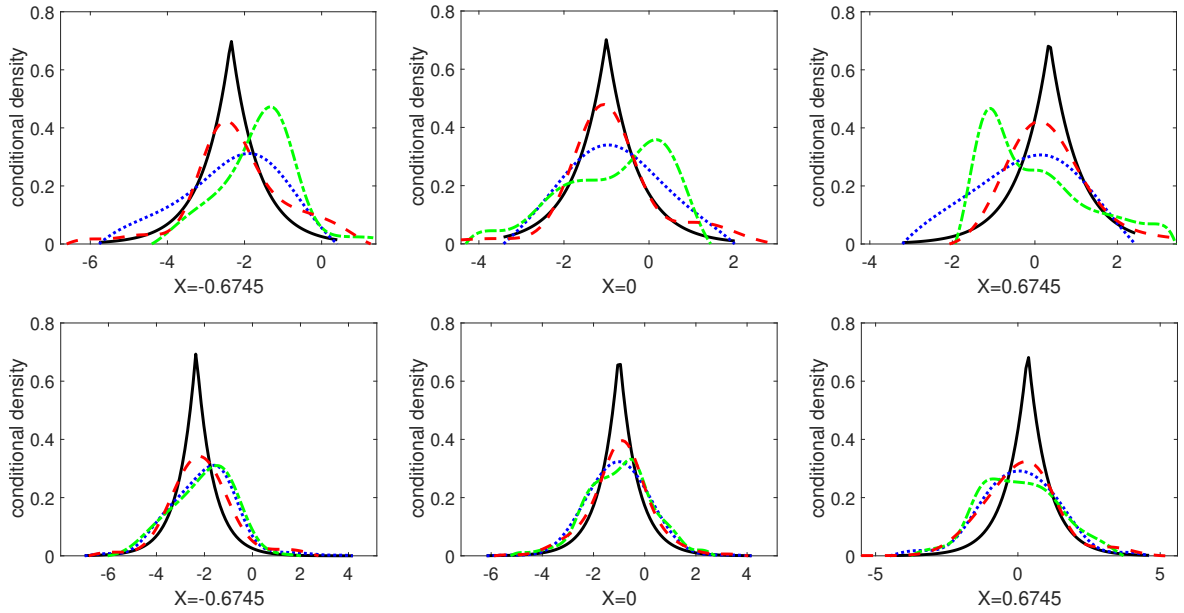

Figure 3: Conditional density at three different locations in the support of the predictors, at sample size $n = 100$ (top row) and $n = 1000$ (bottom row), for a unimodal conditional density.

estimation. However, the bias is substantially reduced for sample size 1000. The average performance based on average $\mathbb{L}^2$ loss function is illustrated in Figure 5 for the two examples at the three locations with sample sizes 100 and 1000. The boxplots indicate that for higher sample size, the average performance and performance stability improve in all cases.

7 Application to DNA methylation profile

As an application of modality-constrained density estimation, we consider the dataset discussed in Eckstein et al. [2017]. The dataset is quite large, with 820374 data points. It is univariate, consisting of methylation levels in HeLa cells, with values between 0 and 1. Values close to zero indicate low methylation levels and values close to 1 indicate uniform (high levels of) methylation in the cell. It is well known in the methylation literature (see the paper by Harris et al. [2010]

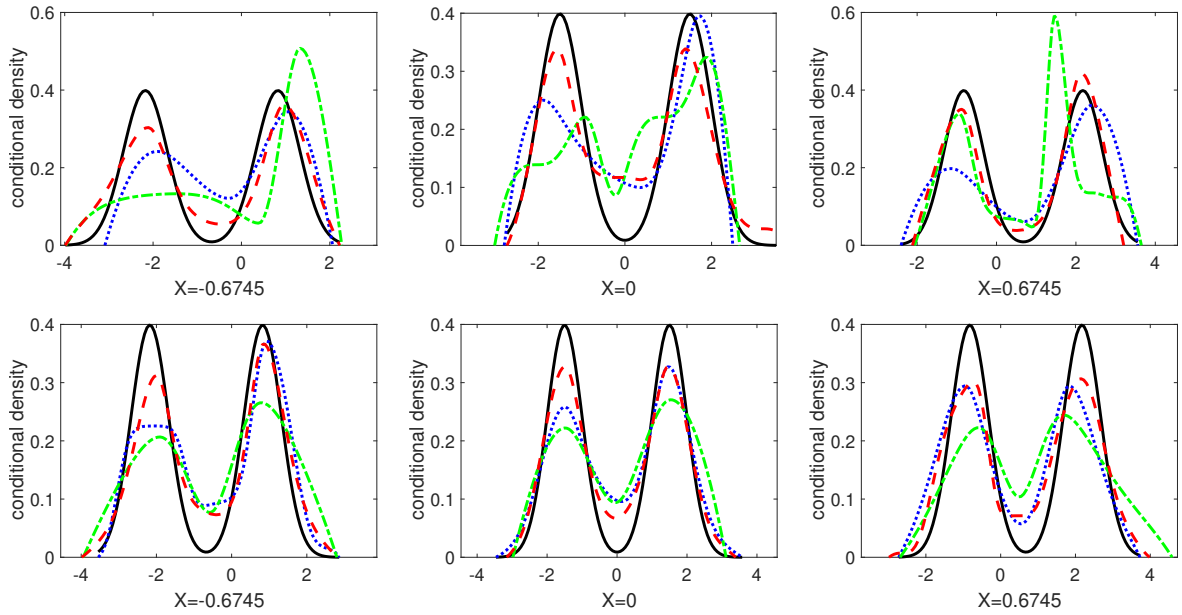

Figure 4: Conditional density at three different locations in the support of the predictors, at sample size $n = 100$ (top row) and $n = 1000$ (bottom row), for a bimodal conditional density.

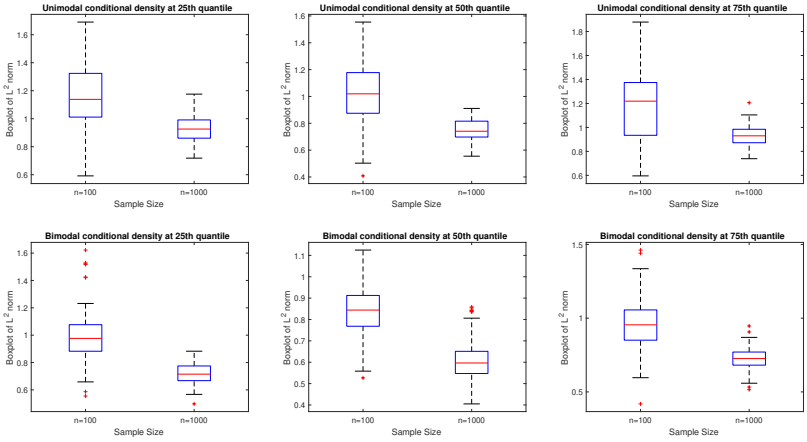

Figure 5: Behaviour of the $\mathbb{L}^2$ norm at different sample sizes and locations for the two conditional density examples.

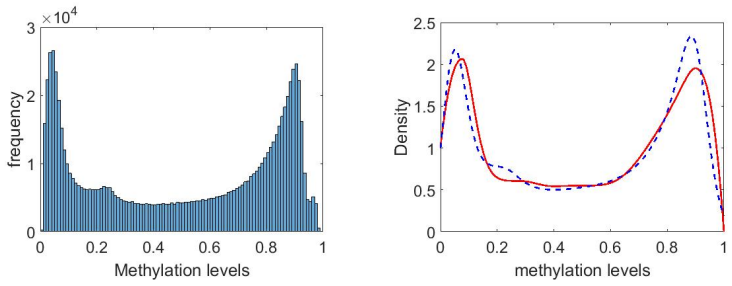

Figure 6: The figure illustrates the histogram (left) and the estimated density of methylation levels (right) for `dtcode` (solid) and `ksdensity` (dashed).

for example), that methylation levels are usually distributed bimodally, because most cells exhibit either very low levels or very high levels. It is thus natural to perform bimodal constrained density estimation.

For this experiment, we use up to 4 basis elements for the tangent space representation of the warping functions. Also, we allow the boundary values of the density estimate to be non-zero. We also present a kernel density estimate found using the inbuilt MATLAB function *ksdensity*, with bandwidth chosen by Silverman's rule of thumb. Figure 6 illustrates the dataset and the performance of the estimators. Note that since the sample size is high, both the estimators give similar performances, though the kernel estimate seems to overestimate the right boundary.

References

Sutanoy Dasgupta, Debdeep Pati, and Anuj Srivastava. A two-step geometric framework for density modeling. *Statistica Sinica*, In press. URL `doi:10.5705/ss.202018.0231`.

Meredith Eckstein, Matthew Rea, and Yvonne N Fondufe-Mittendorf. Microarray dataset of

transient and permanent dna methylation changes in hela cells undergoing inorganic arsenic-mediated epithelial-to-mesenchymal transition. *Data in brief*, 13:6–9, 2017.

R Alan Harris, Ting Wang, Cristian Coarfa, Raman P Nagarajan, Chibo Hong, Sara L Downey, Brett E Johnson, Shaun D Fouse, Allen Delaney, Yongjun Zhao, et al. Comparison of sequencing-based methods to profile dna methylation and identification of monoallelic epigenetic modifications. *Nature biotechnology*, 28(10):1097, 2010.

Hans Triebel. Theory of function spaces. iii, volume 100 of monographs in mathematics. *BirkhauserVerlag, Basel*, 2006.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, pages 339–362, 1995.