Modality-Constrained Density Estimation via Deformable Templates

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 paper 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. The supplementary materials for our paper can be found online.

Keywords: Density estimation, Modality constraints, Shape constraints, Sieve estimation, Deformation group, Conditional densities,

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 in order 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 papers 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 papers 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 papers assumed knowledge of the position and value of the mode, and applied monotonic estimators over subintervals on either side of it; later papers, for example 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 provide 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. (2009) 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: *e.g.* see the very recent article by Wheeler et al. (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 paper Cheng et al. (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 et al. (ress) introduced a geometric approach for exploring the space of *all* probability densities in order 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 paper, we take a principled and geometrically-intuitive maximum-likelihood approach to the problem of modality-constrained density estimation. The primary contribution of this paper 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 et al. (ress). 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 et al. (1999), we apply only a single transformation rather than a possibly non-convergent sequence. Coupled with a small number of other parameters, this transformation constitutes a parametrization 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 et al. (ress) is in the choice of transformations used. Dasgupta et al. (ress) wish to parameterize the set of all positive densities. As a result they choose a set of transformations that act transitively, *i.e.* 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, *i.e.* a one-to-one, differentiable map whose inverse is also differentiable. Here, $\dot{\gamma}$ is the derivative of γ, that is, $\dot{\gamma}(t) = \dfrac{\mathrm{d}\gamma(t)}{\mathrm{d}t}$ 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. In order 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 $\mathcal{P}_M = \{p : [0,1] \to \mathbb{R}_+ :$

$p(0) = p(1) = 0, p$ has M interior modes$\}$.

- Let the critical points of a pdf $p \in \mathcal{P}_M$ with M modes be located at $\{b_a : a \in \{0, \cdots, 2M\}\}$, 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 $(a+1)^{\text{th}}$ 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 $\underset{\lambda}{\cup}\mathcal{P}_{M,\lambda}$ 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, *i.e.* 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, *i.e.* 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.

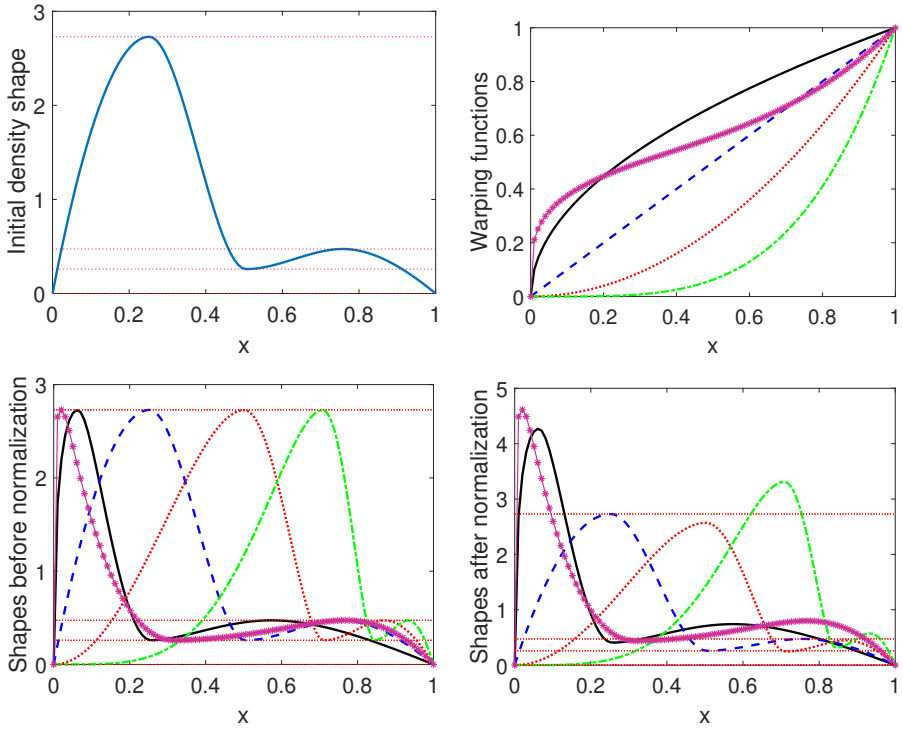

Figure 1: Top left: Initial density. Top right: Different warping functions. Bottom left: Shapes resulting from warping the initial density without renormalization. Bottom right: Resultant warped densities after renormalization.

Figure 1 illustrates the height-ratio-vector-preserving effect of the transformations by applying several elements of Γ to a 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 j^{th} 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 $M > 1$, $\Lambda_M = \{\lambda \in \mathbb{R}_+^{(2M-2)} : \lambda_1 < 1, \lambda_1 < \lambda_2, \lambda_{2j+1} < \lambda_{2j}, \lambda_{2j+1} < \lambda_{2j+2}, j = 1, 2, \cdots, M-2\}$, 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.

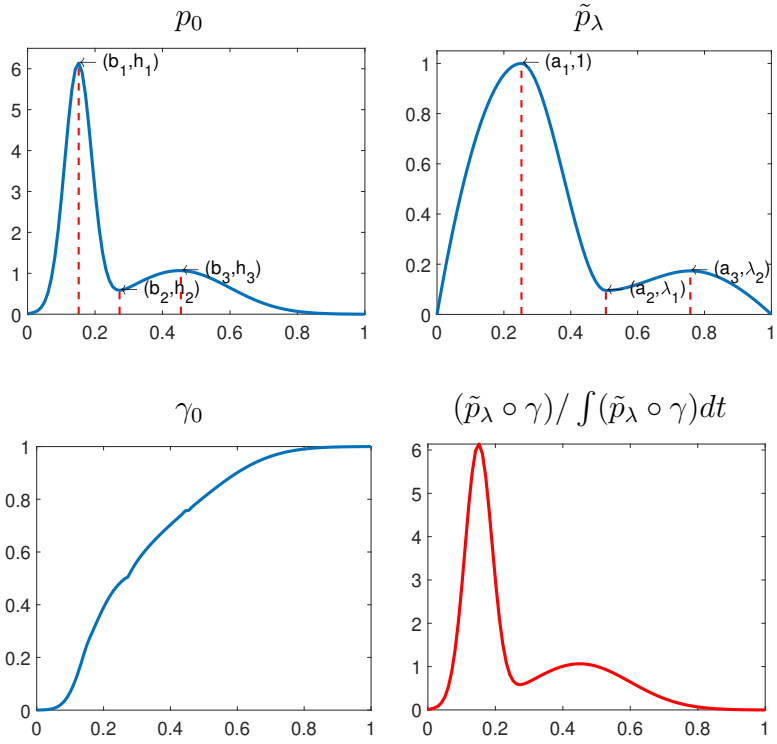

Figure 2: Top left: The original density. Top right: Initial template. Bottom left: The γ_0 transforming the template to original shape. Bottom right: Reconstructed density.

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 paper, 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, *i.e.* 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 $\gamma = \gamma_{\text{id}}(t) = t$, *i.e.* 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:

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

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 $V = \{v \in T_{\mathbf{1}}(\mathbb{S}_\infty) : \|v\| < \pi\}$.

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

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

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) discusses 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 $\mathcal{C} = \{c = (c_1, c_2, \cdots) : \sum_{j=1}^{\infty} c_j b_j(t) \in V\}$. 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

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

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

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

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 2 basis elements, and increase the number up to a pre-decided 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 overpenalizes 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 seconds for a sample of size 100, to 250 seconds 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.

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 $p_0 = 0.95\mathcal{N}(0,0.5) + 0.05\mathcal{N}(3,1)$.

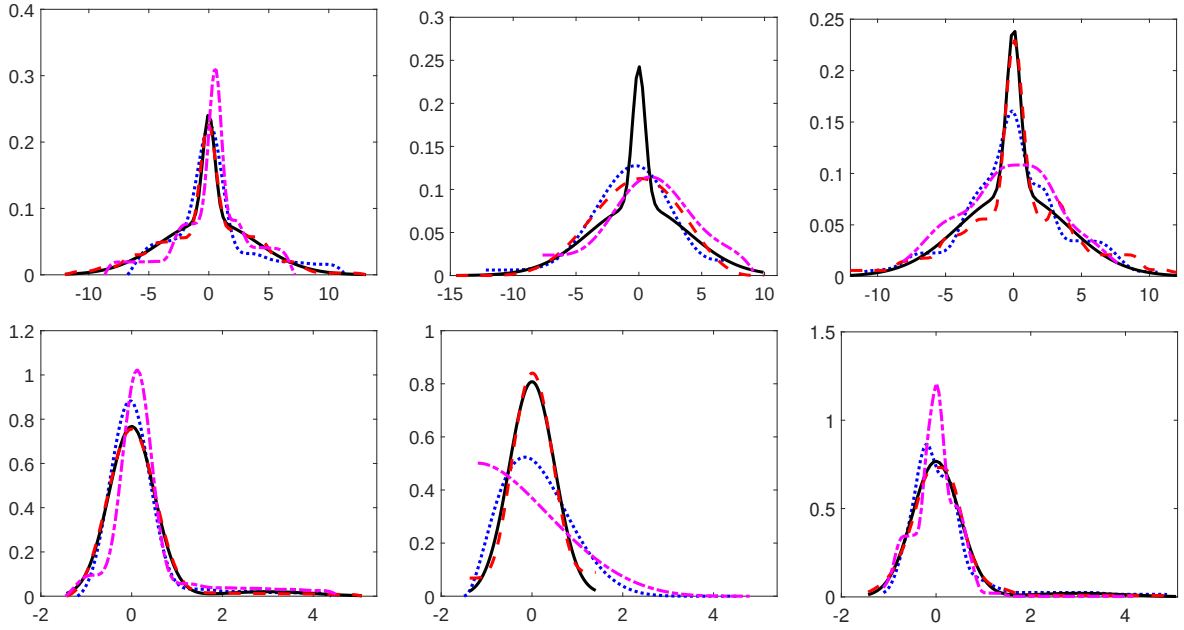

Figure 3: Performance of `dtcode` (left) vs. `umd` (centre) and `scdensity` (right) for examples 1 (top row) and 2 (bottom row) of Study 1, using the $\mathbb{L}^2$ loss function values calculated for 100 samples of size 100. The true density is shown as a solid line; the estimated density with best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

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) (centre 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 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

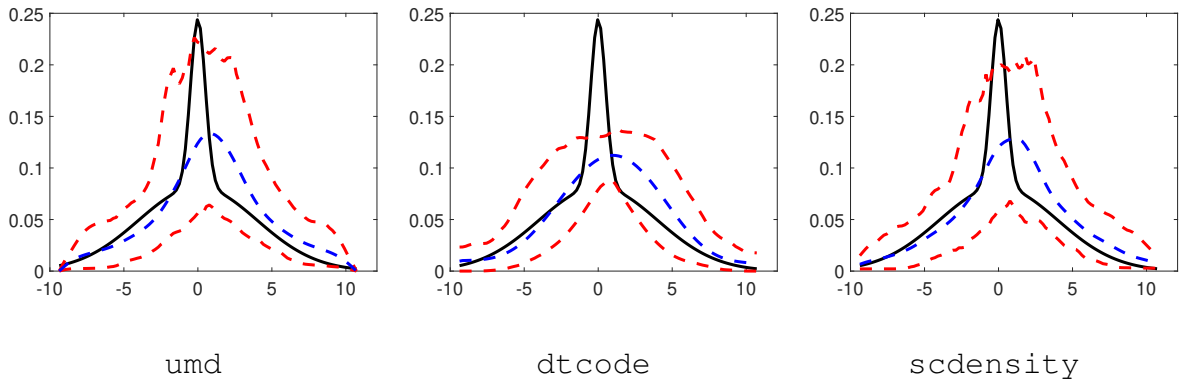

Figure 4: A comparison of the variability of the estimates across different samples for example 1 from Study 1 at sample size 100. The middle dashed line indicates the average estimate across samples, while the upper and lower dashed lines represent the 95^{th} and 5^{th} quantiles, respectively, of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

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.

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 pre-specified 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. Secondly, 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 in the original paper 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 pre-specified 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

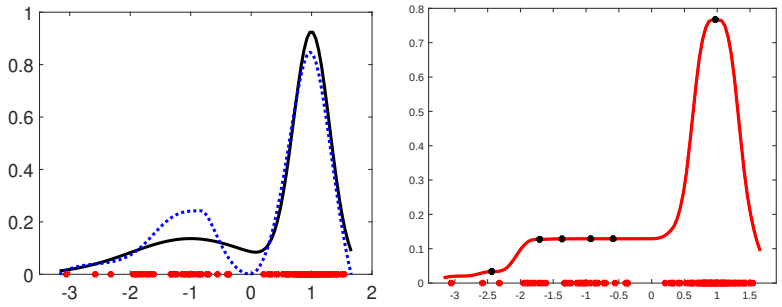

Figure 5: The left panel shows the bimodal example from Study 2, with the true density (solid) and `dtcode` estimate(dotted). The right panel shows the `scdensity` estimate along with local modes. The observations are shown on the x-axis.

to more general modality constraints.

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)

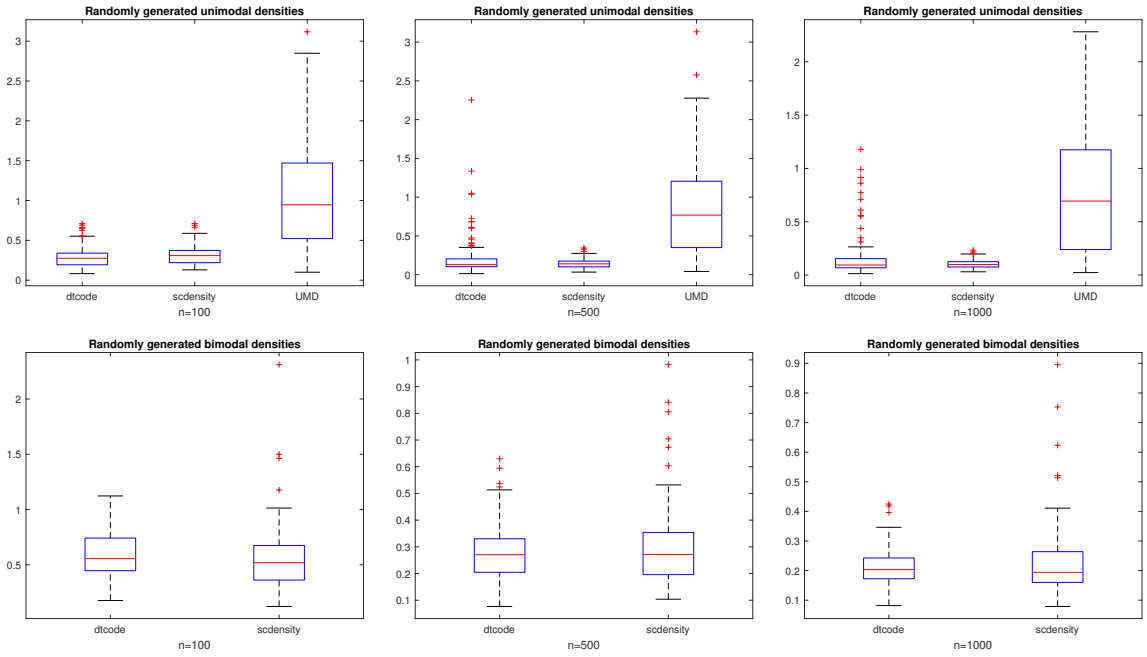

Figure 6: Boxplots of the $\mathbb{L}^2$ norms of the errors of the density estimates found using `dtcode`, `scdensity`, and `umd`, for randomly sampled true densities and different sample sizes. The top row shows the results for example 1; the bootom row those for example 2.

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.

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 $p_0 = 0.25\mathcal{N}(-4, 0.5) + 0.25\mathcal{N}(-2, 0.5) +$

$$0.4\mathcal{N}(2,1) + 0.1\mathcal{N}(5, 0.25).$$

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 et al. (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,

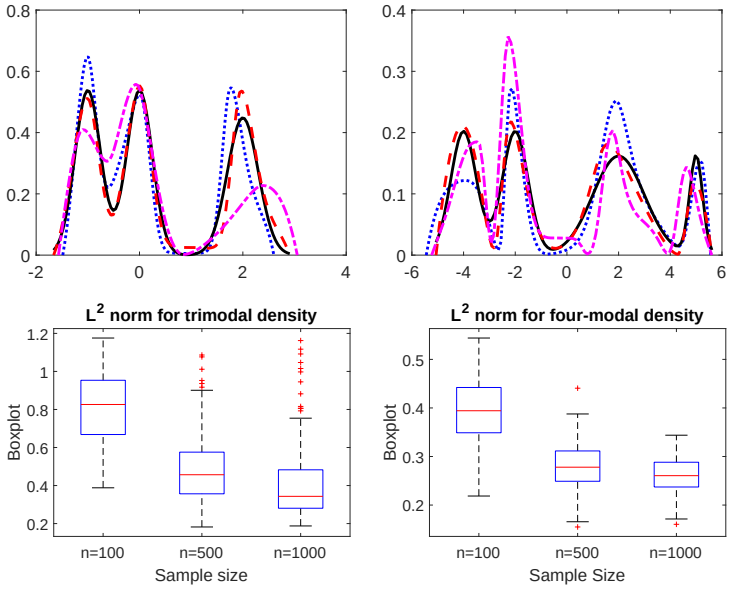

Figure 7: Results of the `dtcode` method on examples 1 (left column) and 2 (right column) of Study 4. The top row shows plots of the true density as a solid line; the estimated density: with best performance as a dashed line; with median performance as a dotted line; and with worst performance as a dashed-dotted line; woith 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.

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, in order 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.

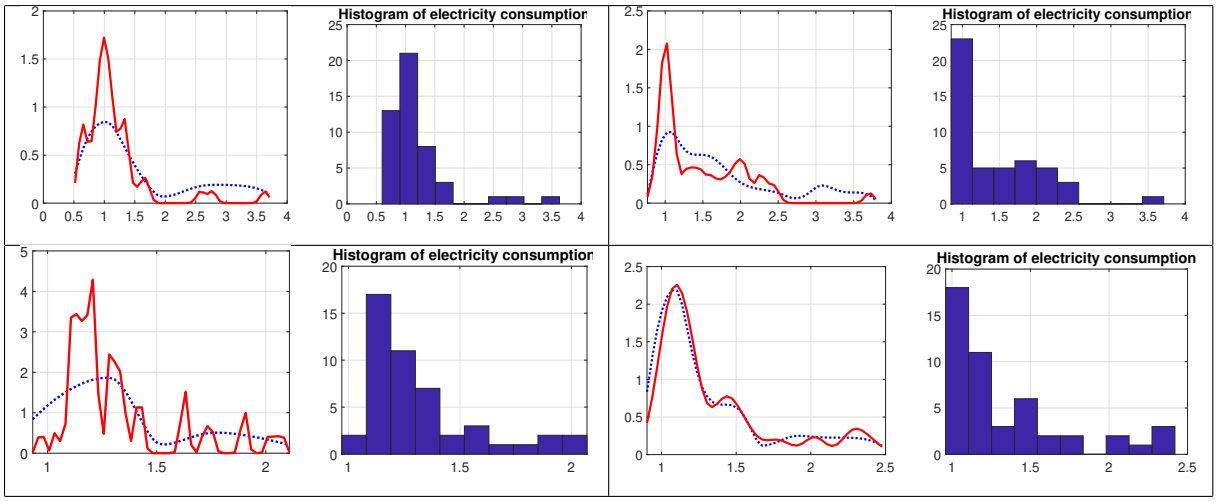

Figure 8: Estimated densities of electricity consumption using the warped approach (dotted) and a kernel (ucv) approach (solid), and the associated histograms of consumption data for four different days.

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 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:

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

where $W_{x_0,i}$ is the localized weight associated with the i^{th} 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 neighbourhood 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 in order 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 et al. (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 distribution for traffic flows between 1000 and 1620 vehicles per lane per hour, corresponding to uncongested and congested traffic. This range of traffic flows has already been studied by Einbeck and Tutz (2006). They note 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 35.56mph and the right mode is at 59.01mph. Einbeck and Tutz (2006) obtain a very similar conditional density estimate. The left mode in their case is at 32.65mph and the right mode is at 59.18mph. 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.

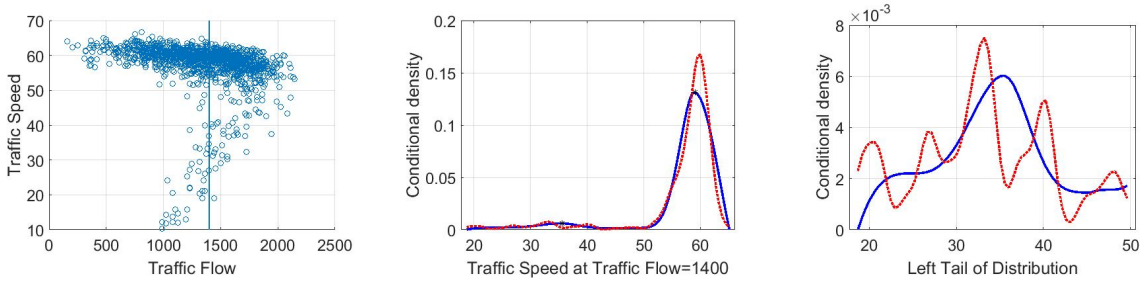

Figure 9: Left: scatterplot of traffic flow data for Californian highways from the `hdrcde` package in R. Centre: traffic speed density at traffic flow 1400 as estimated by `dtcode` (solid line) and the `NP` package (dotted line). Right: a magnified view of the left part of the centre plot.

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 (greater than 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 paper 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 paper 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 paper, 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 paper 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.

9 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 paper. 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.

References

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

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

Bickel, P. J. and J. Fan (1996). Some problems on the estimation of unimodal densities. *Stat. Sin. 6*(1), 23–45.

Birge, L. (1997). Estimation of unimodal densities without smoothness assumptions. *Ann. Stat. 25*(3), 970–981.

Brunner, L. J. and A. Y. Lo (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Stat. 17*(4), 1550–1566.

Cheng, M.-Y., T. Gasser, and P. Hall (1999). Nonparametric density estimation under unimodality and monotonicity constraints. *J. Comput. Graph. Stat. 8*(1), 1–21.

Cordova, J., L. M. K. Sriram, A. Kocatepe, Y. Zhou, E. E. Ozguven, and R. Arghandeh (2018). Combined electricity and traffic short-term load forecasting using bundled causality engine. *IEEE Transactions on Intelligent Transportation Systems*, 1–11.

Dasgupta, S., D. Pati, and A. Srivastava (In press). A two-step geometric framework for density modeling. *Statistica Sinica*.

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

Einbeck, J. and G. Tutz (2006). Modelling beyond regression functions: an application of multimodal regression to speed–flow data. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 55*(4), 461–475.

Grenander, U. (1956). On the theory of mortality measurement: part ii. *Scand. Actuar. J.*.

Hall, P. and L.-S. Huang (2002). Unimodal density estimation using kernel methods. *Statistica Sinica*, 965–990.

Harris, R. A., T. Wang, C. Coarfa, R. P. Nagarajan, C. Hong, S. L. Downey, B. E. Johnson, S. D. Fouse, A. Delaney, Y. Zhao, et al. (2010). Comparison of sequencing-based methods to profile dna methylation and identification of monoallelic epigenetic modifications. *Nature biotechnology 28*(10), 1097.

Izenman, A. J. (1991). Review papers: Recent developments in nonparametric density estimation. *J. Am. Stat. Assoc. 86*(413), 205–224.

Kwac, J., J. Flora, and R. Rajagopal (2014, Jan). Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid 5*(1), 420–430.

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

Meyer, M. C. (2001). An alternative unimodal density estimator with a consistent estimate of the mode. *Stat. Sin. 11*(4), 1159–1174.

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom, and A. Siepel (2009). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*.

Rao, B. L. S. P. (1969). Estimation of a unimodal density. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) 31*(1), 23–36.

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

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

Turnbull, B. C. and S. K. Ghosh (2014). Unimodal density estimation using bernstein polynomials. *Comput. Stat. Data Anal. 72*, 13–29.

Ugray, Z., L. Lasdon, J. Plummer, F. Glover, J. Kelly, and R. Martí (2007). Scatter search and local nlp solvers: A multistart framework for global optimization. *INFORMS Journal on Computing 19*(3), 328–340.

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

Wahba, G. (1981). Data-based optimal smoothing of orthogonal series density estimates. *The annals of statistics*, 146–156.

Wasserman, L. and J. D. Lafferty (2006). Rodeo: Sparse nonparametric regression in high dimensions. In *Advances in Neural Information Processing Systems*, pp. 707–714.

Wegman, E. J. (1970). Maximum likelihood estimation of a unimodal density, II. *Ann. Math. Stat. 41*(6), 2169–2174.

Wheeler, M., D. Dunson, and A. Herring (2017). Bayesian local extremum splines. *Biometrika 104*(4), 939–952.

Wolters, M. A. and W. J. Braun (2018). A practical implementation of weighted kernel density estimation for handling shape constraints. *Stat 7*(1), e202.