SHAPE-CONSTRAINED AND UNCONSTRAINED DENSITY ESTIMATION USING GEOMETRIC EXPLORATION

Sutanoy Dasgupta, Debdeep Pati, Ian Jermyn and Anuj Srivastava

Florida State University

ABSTRACT

The problem of nonparametrically estimating probability density functions (*pdf*s) from observed data requires posing and solving optimization problems on the space of *pdf*s. We take a geometric approach and explore this space for optimization using actions of a time-warping group. One action, termed *area preserving*, is transitive and is applicable to the case of unconstrained density estimation. In this case, we take a two-step approach that involves obtaining any initial estimate of the *pdf* and then transforming it via this warping function to reach the final estimate, while maximizing the log-likelihood function. Another action, termed *mode-preserving*, is useful in situations where the *pdf* is constrained in shape, i.e. the number of its modes is known. As earlier, we initialize the estimation with an arbitrary element of the correct shape class, and then search over all time warpings to reach the optimal *pdf* within that shape class. Optimization over warping functions is performed numerically using the geometry of the group of warping functions. These methods are illustrated using a number of simulated examples.

Index Terms— Density estimation, time warping, Shape-constrained density, optimization on sphere.

1. INTRODUCTION

Estimating probability density functions (*pdf*s) from sampled data is an important and well-studied field of research in statistical inference. The most basic problem in this area is that of estimating a univariate *pdf* from its *iid* samples [1, 2, 3, 4, 5, 6, 7]. The problem gets more challenging when one imposes additional constraints on the estimate, especially those on the shape of the *pdf*s allowed. Imposition of such constraints is motivated by the fact that if the true density is known to have a certain shape class, say unimodal, bimodal or trimodal, then one should be able to leverage that knowledge into improving estimation accuracy. Also ensuring that the estimate has the correct number of modes improves the usefulness of the estimate as an exploratory tool. The difficulty also grows as one goes to multivariate density estimation [8, 9, 11].

A majority of solutions in density estimation boil down to formulating and optimizing an objective function over $\mathscr{F}$, the space of all *pdf*s. To keep the discussion simple, we will assume that $\mathscr{F}$ is the space of all densities strictly positive on the domain and zero elsewhere. Also, although the proposed methodology applies to higher-dimensional domains (see a related paper in 2D [14]), we will restrict to the univariate case here. The objective functions for estimation can come from different motivations, but the search space for an optimal solution remains to be $\mathscr{F}$. Most of the past literature has focused on the nature and formulation of these objective functions – frequentist or Bayesian, parametric or nonparametric, penalized or non-penalized – but the focus is seldom on the nature of the space being searched [15]. We follow the logic that if there is an efficient way to explore $\mathscr{F}$, then the associated optimization solution becomes efficient also. Taking a frequentist, nonparametric approach, we will handle the geometry of $\mathscr{F}$ using actions of a time warping group. Let Γ be the set of all positive diffeomorphisms from $[0, 1]$ to itself, i.e. $\Gamma = \{\gamma | \gamma \text{ is differentiable}, \gamma^{-1} \text{ is differentiable}, \dot{\gamma} > 0, \gamma(0) = 0, \gamma(1) = 1\}$. The elements of Γ play the role of warping functions, or transformations of *pdf*s. There are several actions possible and we will use two in the paper for exploring $\mathscr{F}$:

1. **Area-Preserving Action**: For any $f \in \mathscr{F}$ and $\gamma \in \Gamma$, the mapping $(f, \gamma) = (f \circ \gamma)\dot{\gamma}$ defines an action that is area preserving.
2. **Mode-Preserving**: For any $f \in \mathscr{F}$ and $\gamma \in \Gamma$, the mapping $(f, \gamma) = \frac{(f \circ \gamma)}{\int (f \circ \gamma) ds}$ defines an action that preserves the number of modes of f.

One can show that both these mappings form proper group actions of Γ on $\mathcal{F}$.

Based on these two actions, we target two estimation scenarios. (1) **Scenario 1**: Here we focus on an unconstrained density estimation, i.e. simple estimation of a *pdf* from the data without any additional constraints, via a two-step process. The first step seeks a computationally fast, albeit sub-optimal density estimate f_p from the data. The second step involves transforming f_p using the area-preserving action of Γ to obtain the final estimate. The second step requires solving an optimization problem on Γ under the chosen criterion (say MLE or penalized-MLE). (2) **Scenario 2**: Here we study a situation where the number of modes in underlying *pdf* are

Thanks to XYZ agency for funding.

known. This problem, called *shape-constrained* density estimation, is quite challenging because ensuring both a correct shape and the optimality of the estimate under the chosen criterion is seemingly complicated. There is no known literature on density estimation with multimodal constraints in the past. Once again we take a two-step approach where the first step constructs an arbitrary *template* that satisfies the given constraints. The second step uses the mode-preserving action of Γ and transforms the template into better solutions. As before, the second step requires solving an optimization problem over Γ under the chosen criterion (MLE or penalized-MLE). Additionally, we search over different heights of the function at the critical points, to reach the full space of desired shapes. This joint search over the two unknowns – time warping and the vector of heights – is performed using a numerical approach. Experimental results demonstrate the success of the proposed framework in both the scenarios.

The rest of this paper is as follows. Section 2 discusses the proposed framework for unconstrained and constrained density estimation, and derives the objective functions. Section 3 presents the estimation procedure by optimizing over the set of warping functions. Section 4 presents some simulation examples and Section 5 ends the paper with a short discussion.

2. METHODOLOGY

In this section we introduce our framework for density estimation in two scenarios: (i) unconstrained and (ii) mode-constrained (or shape-constrained) estimation. In each case, the estimation procedure involves making an initial guess (from the correct constraint class when needed) and warping it optimally to find a final estimator. We setup these estimation problems first and focus on the optimization procedure later.

2.1. Unconstrained Density Estimation

As the first problem, we focus on the problem of estimating a univariate *pdf* (call it f_0) from its *iid* samples. For simplicity of exposition, we will assume that $f_0 > 0$ although this condition can be easily relaxed. Since we need to explore the full *pdf* space, we will use the *area-preserving* action of Γ on $\mathscr{F}$. Let $f_p \in \mathscr{F}$ be an initial estimate; this estimate can be obtained using a parametric assumption or any other fast nonparametric estimate. In principle, any element of $\mathscr{F}$ is sufficient for the purpose but in practice the idea is to get close to f_0 while being computationally efficient. Once we have an initial guess f_p, we plug the gap between f_p and the optimal solution using the action: $(f_p, \gamma) \to \hat{f} \equiv (f_p \circ \gamma)\dot{\gamma}$. This action is called **area-preserving** because $\int_0^1 \hat{f}(x)dx = \int_0^1 f_p(\gamma(x))\dot{\gamma}(x)dx = \int_0^1 f_p(x)dx$. In other words, a pdf f_p remains a *pdf* under this transformation. Furthermore, this action is transitive. That is, one can go from any element of $\mathscr{F}$ to any other element of $\mathscr{F}$ using a unique element of Γ. This property makes this framework very powerful; this implies that *any* initial guess f_p is sufficient for this search.

What should be the criterion for optimization? Taking the MLE approach, we seek an estimate $\hat{f}$ the maximizes the log-likelihood of the given data. This sets up an optimization problem over Γ. Given sample observations $\{x_i, i = 1, 2, \cdots, n\}$ from f_0 and an initial density estimate $f_p \in \mathscr{F}$, the final estimate is given by $\hat{f}(t) = f_p(\hat{\gamma}(t))\dot{\hat{\gamma}}(t), t \in [0, 1]$, where

$$\hat{\gamma} = \underset{\gamma \in \Gamma}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \left[\log(f_p(\gamma(x_i))) + \log(\dot{\gamma}(x_i)) \right] \right) .$$

The quantity in the parenthesis is exactly the log-likelihood of the given data under the estimated density. One can also add a regularization term involving either the estimated density $\hat{f}$ or the time-warping function γ, if needed. More generally, one can replace this cost with a Bayesian cost function also.

The next issue is the optimization over Γ. This problem is complicated because Γ is an infinite-dimensional, nonlinear manifold. As described later in Section 3, we use a combination of local flattening and a truncated basis expansion to represent elements of (a large subset) of Γ via finite-dimensional vectors $c \in \mathbb{R}^d$. Thus, we can optimize over this Euclidean space using standard optimization tools in matlab. Let for any $c \in \mathbb{R}^d$, $\gamma_c \in \Gamma$ denote the corresponding warping function (see Section 3 for details). Then, the final solution $\hat{f}$ uses $\hat{\gamma} = \gamma_{\hat{c}}$, where

$$\hat{c} = \underset{c \in \mathbb{R}^d}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \left[\log\left(f_p\left(\gamma_c(x_i)\right) \dot{\gamma}_c(x_i)\right) \right] \right).$$

All that remains is to solve this optimization problem and one can use any convenient numerical tool for that purpose.

2.2. Mode- or Shape-Constrained Density Estimation

Next we consider the problem of estimating a *pdf* from data in situation where the number of modes are pre-specified. While there has been past work on unimodal or log-concave density estimation [10, 12, 13], there is little work on the problem of multimodal density estimation. We will take the time-warping approach as earlier, but this time we use the mode-preserving action of Γ on $\mathscr{F}$. We point out that while the number of modes is given, the heights of the function at modes, or at the critical points, are not specified. One has to search over both the placements and the heights of the critical points in order to reach an optimal estimate.

To setup this estimation problem, we introduce some extra notation. Let $f \in \mathscr{F}$ be a *pdf* with m well-defined modes, and let the critical points of f be located at $b_i \in [0, 1]$, $i = 0, \cdots, 2m$ with $b_0 = 0$ and $b_1 = 1$. We define the *height-ratio vector* of f to be $\lambda^f = (\lambda_1, \lambda_2, \ldots, \lambda_{2m-2})$, where $\lambda_i = f(b_{i+1})/f(b_1)$ is the ratio of the height of the $(i+1)^{st}$ interior critical point to the height of the first (from the left) mode. Let h_1 be the (unknown) height of the left most mode of f.

Consider the action of Γ on $\mathscr{F}$ given by the mapping $(f, \gamma) = \tilde{f} \equiv \frac{f \circ \gamma}{\int (f \circ \gamma) dx}$. It is interesting to note that under this action: (i) the number of modes of $\tilde{f}$ is same as that of f, only the locations are changed, and (ii) the height-ratio vector of $\tilde{f}$ remains same as that of f, i.e. $\lambda^{\tilde{f}} = \lambda^f$. Then the estimation process is as follows.

Template: Construct any pdf g with m well-defined modes. One way to do this is to construct a g with conditions: $g(0) = g(1) = 0$; the locations of the intermediate critical points are uniformly spaced in $(0, 1)$, with $b_0 = 0$, and $b_{2M} = 1$; and $g(b_1) = 1$. Let $\Lambda = \{\lambda \in \mathbb{R}^{(2M-2)^+} | \lambda_1 < 1, \lambda_1 < \lambda_2, \lambda_{2j+1} < \lambda_{2j}, \lambda_{2j+1} < \lambda_{2j+2}, j = 1, 2, \cdots, M-2\}$. Choose an arbitrary height-ratio vector $\lambda \in \Lambda$, and set the heights of g at b_is such that the height-ratio vector of g is this λ. Obtain g at the remaining points through linear interpolation. Call this g as g_λ.

Optimization: Given such a g_λ our estimate is given by $\frac{g_\lambda \circ \gamma}{\int_0^1 (g_\lambda \circ \gamma)\, dt}$ where $\gamma \in \Gamma = \{\gamma : [0,1] \to [0,1] | \dot{\gamma} > 0, \gamma(0) = 0, \gamma(1) = 1\}$. Thus, the two variable of interest for optimization are γ and λ. The maximum likelihood estimate of the underlying density, given the initial template function $g = g_\lambda$, is $\hat{f}(t) = g_{\hat{\lambda}}(\gamma_{\hat{c}}(t)) / (\int_0^1 g_{\hat{\lambda}}(\gamma_{\hat{c}}(t)) dt), t \in [0, 1]$, where $\gamma_{\hat{c}} = H^{-1}(c)$ defined in the next section, and

$$(\hat{c}, \hat{\lambda}) = \underset{c \in \mathbb{R}^d, \lambda \in \Lambda}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \left[\log \left(\frac{g_\lambda\left(\gamma_c(X_i)\right)}{\int_0^1 \left(g_\lambda\left(\gamma_c(t)\right) dt\right)} \right) \right] \right). \tag{1}$$

Once again, all that remains is solving this joint optimization problem, and accomplish this using numerical tools.

3. OPTIMIZATION OVER WARPINGS GROUP

The proposed framework for density estimation, both the shape-constrained and unconstrained cases, leads to a certain optimization problem on Γ. This optimization is made challenging by the fact that Γ is an infinite dimensional manifold We handle the nonlinearity by forming a bijective map from Γ to a tangent space of the unit Hilbert sphere $\mathbb{S}_\infty$ (the tangent space is a vector space), and infinite dimensionality by selecting a finite-dimensional subspace of this tangent space. Together, these two steps are equivalent to finding a family of finite-dimensional submanifolds of Γ that can be *flattened* into vector spaces. This, in turn, allows for representing any γ using elements of a Euclidean vector space and use of standard optimization procedures.

To locally flatten Γ, we define a function $q : [0,1] \to \mathbb{R}$, $q(t) = \sqrt{\dot{\gamma}(t)}$, 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 [16]). For any $\gamma \in \Gamma$, its SRSF q is an element of the positive orthant of the unit Hilbert sphere $\mathbb{S}_\infty \subset \mathbb{L}^2$, denoted by $\mathbb{S}_\infty^+$. The set $\mathbb{S}_\infty$ is a smooth manifold with known geometry under the $\mathbb{L}^2$ Riemannian metric [17]. Although it is not a vector space, it can be easily flattened into a vector space (locally) due to its constant curvature. A natural choice for flattening is the vector space tangent to $\mathbb{S}_\infty^+$ at the point $\mathbf{1}$, which is a constant function with value 1. ($\mathbf{1}$ is the SRSF corresponding to $\gamma = \gamma_{\mathrm{id}}(t) = t$.) The tangent space of $\mathbb{S}_\infty^+$ at $\mathbf{1}$ is an infinite-dimensional vector space given by: $T_{\mathbf{1}}(\mathbb{S}_\infty^+) = \{v \in \mathbb{L}^2([0,1], \mathbb{R}) | \int_0^1 v(t) dt = \langle v, \mathbf{1} \rangle = 0\}$. Next, we define a mapping that takes an arbitrary element of $\mathbb{S}_\infty^+$ to this tangent space. For this *retraction*, we will use the inverse exponential map which takes any $q \in \mathbb{S}_\infty^+$ to $T_{\mathbf{1}}(\mathbb{S}_\infty^+)$ according to:

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

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: $\langle v_1, v_2 \rangle = \int_0^1 v_1(t) v_2(t) dt$. Further, we can select any orthogonal basis $\mathcal{B} = \{b_j, j = 1, 2, \ldots\}$ of the Hilbert space $T_{\mathbf{1}}(\mathbb{S}_\infty^+)$ to express its elements v by their corresponding coefficients; that is, $v(t) = \sum_{j=1}^\infty c_j b_j(t)$, where $c_j = \langle v, b_j \rangle$. The only restriction on the basis elements b_j's is that they must be orthogonal to $\mathbf{1}$, that is, $\langle b_j, \mathbf{1} \rangle = 0$. In order to map points back from the tangent space to the Hilbert sphere, we use the exponential map, given by:

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

In practice, we restrict the range and the domain of the exponential map (and its inverse) to be able go back and forth between $\mathbb{S}_\infty^+$ and $T_{\mathbf{1}}(\mathbb{S}_\infty^+)$. Using these two steps, we specify the finite-dimensional, and therefore approximate, representation of warpings. We define a composite map $H : \Gamma \to \mathbb{R}^J$, illustrated in Figure 1, as follows.

$$\gamma \in \Gamma \xrightarrow{\sqrt{\dot{\gamma}}} \mathbb{S}_\infty^+ \xrightarrow{\exp_{\mathbf{1}}^{-1}} v \in T_{\mathbf{1}}^0(\mathbb{S}_\infty^+) \xrightarrow{\{b_j\}} \{c_j = \langle v, b_j \rangle\}. \tag{4}$$

Let $V_\pi^J = \{c \in \mathbb{R}^J : \| \sum_{j=1}^J c_j b_j \| < \pi/4\} \subset \mathbb{R}^J$. For any $c \in V_\pi^J$, let γ_c denote the diffeomorphism $H^{-1}(c)$. For any fixed J, the set $H^{-1}(V_\pi^J)$ is a J-dimensional submanifold of Γ,and we pose the estimation problem on this submanifold. As J goes to infinity, this submanifold converges to the full group Γ.

With this representation, any optimization problem on Γ can be transferred to the set V_π^J using H and its inverse. We solve this Euclidean optimization problem using function `fminsearch` in matlab. One can make the choice of J adaptive to data but that is left for future work.

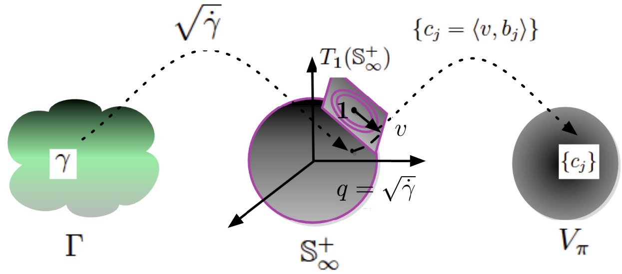

Fig. 1. Finite dimensional representation of Γ .

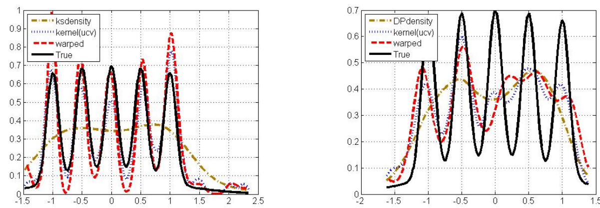

Fig. 3. *The left panel shows the improvement over initial ksdensity estimate. Both kernel(ucv) and warped estimate have a good performance here. The right panel shows that all the methods fail to capture all the peaks. Kernel(ucv) performance is very similar to the warped estimate.*

4. EXPERIMENTAL RESULTS

In this section we present some illustrative experimental results on estimating *pdf*s in the two scenarios laid out earlier.

4.1. Unconstrained Density Estimation

For Scenario 1, we present an illustration using two examples with true underlying densities being: (1) $f_0 \propto 0.75 \exp 3 + 0.25\mathcal{N}(0.75, 2^2)$, truncated to the unit interval $[0, 1]$ (shown in Fig. 2), and (2) $f_0 = \frac{1}{2}\mathcal{N}(0, 1) + \sum_{l=0}^{4} \frac{1}{10}\mathcal{N}(\frac{l}{2} - 1, (0.1)^2)$, a claw density shown in Fig. 3. We generate $n = 100$ independent samples and apply our framework for density estimation with the initial guess coming from a Gaussian family. For comparison, we use a standard kernel estimate (*kernel(ucv)*) and a Bayesian estimate (*DPDensity*). As these two figures show, our estimates provide better estimates than these state of the art estimators.

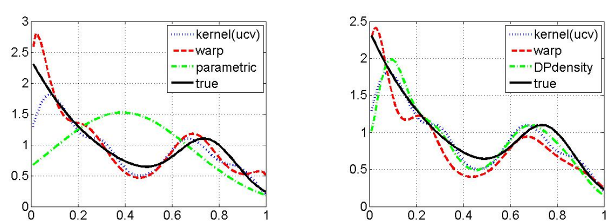

Fig. 2. *The left panel compares the warped estimate $\hat{f}$ with other estimates when f_p is parametric. The right figure compares the warped estimate with others when f_p is ksdensity.*

4.2. Mode- or Shape-Constrained Density Estimation

For Scenario 2, we assume that the number of modes in the underlying density is known. We present some experimental results using unimodal, bimodal, trimodal and monotonic densities. We take 100 simulations of sample size $n = 100$ each (except the monotone example, where we take sample size 500) and present the best, median and worst performance from these simulations in Fig. 4. The examples from top left to bottom right are for densities: (1) $f_0 = 4/5\mathcal{N}(0, 4) + 1/5\mathcal{N}(0, 0.5)$ - a symmetric unimodal example; (2) $f_0 = 1/3\mathcal{N}(-1, 1) + 2/3\mathcal{N}(1, 0.3)$- a asymmetric bimodal example; (3) $f_0 = 1/3\mathcal{N}(-1, 0.25) + 1/3\mathcal{N}(0, 0.25) + 1/3\mathcal{N}(2, 0.3)$ - a asymmetric trimodal example; and, (4) $\mathcal{N}(0, 0.4)I_{[0,1]}$, a monotonically decreasing example. These results underscore the success of our method.

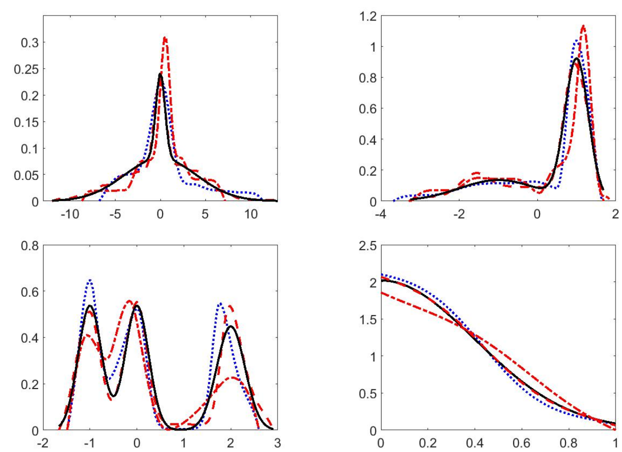

Fig. 4. *The figures show the true density (solid), best (dashed), median (dotted) and worst (dashed-dotted) performance among the 100 samples.*

5. CONCLUSION

The paper presents a geometric approach to density estimation in two specific scenarios. The basic idea is to use the actions of the time warping group to explore the space of pdfs and find the MLE. We introduce two groups actions, one for each scenario, each leading to an optimization problem on the warping group. We solve these optimization problems using the geometry of the warping group and posing a corresponding problem in a finite-dimensional Euclidean space.

References

[1] Qi Li and Jeffrey Scott Racine, *Nonparametric econometrics: theory and practice*, Princeton University Press, 2007.

[2] Tom Leonard, "Density estimation, stochastic processes and prior information," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 113–146, 1978.

[3] Peter J Lenk, "The logistic normal distribution for bayesian, nonparametric, predictive densities," *Journal of the American Statistical Association*, vol. 83, no. 402, pp. 509–516, 1988.

[4] Peter J Lenk, "Towards a practicable bayesian nonparametric density estimator," *Biometrika*, vol. 78, no. 3, pp. 531–543, 1991.

[5] Surya T Tokdar, Yu M Zhu, Jayanta K Ghosh, et al., "Bayesian density regression with logistic gaussian process and subspace projection," *Bayesian analysis*, vol. 5, no. 2, pp. 319–344, 2010.

[6] EG Tabak and Cristina V Turner, "A family of nonparametric density estimation algorithms," *Communications on Pure and Applied Mathematics*, vol. 66, no. 2, pp. 145–164, 2013.

[7] Bruce E Hansen, "Nonparametric conditional density estimation," *Unpublished manuscript*, 2004.

[8] Ulf Grenander, "On the theory of mortality measurement: part ii," *Scandinavian Actuarial Journal*, vol. 1956, no. 2, pp. 125–153, 1956.

[9] Fadoua Balabdaoui and Jon A Wellner, "Estimation of a k-monotone density: limit distribution theory and the spline connection," *The Annals of Statistics*, pp. 2536–2564, 2007.

[10] Arlene KH Kim, Richard J Samworth, et al., "Global rates of convergence in log-concave density estimation," *The Annals of Statistics*, vol. 44, no. 6, pp. 2756–2779, 2016.

[11] Madeleine Cule, Richard Samworth, and Michael Stewart, "Maximum likelihood estimation of a multi-dimensional log-concave density," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 5, pp. 545–607, 2010.

[12] Peter Hall and Li-Shan Huang, "Unimodal density estimation using kernel methods," *Statistica Sinica*, pp. 965–990, 2002.

[13] Bradley C Turnbull and Sujit K Ghosh, "Unimodal density estimation using bernstein polynomials," *Computational Statistics & Data Analysis*, vol. 72, pp. 13–29, 2014.

[14] Ethan Anderes and Marc A Coram, "Two-dimensional density estimation using smooth invertible transformations," *Journal of Statistical Planning and Inference*, vol. 141, no. 3, pp. 1183–1193, 2011.

[15] Torsten Hothorn, Lisa Möst, and Peter Bühlmann, "Most likely transformations," *arXiv preprint arXiv:1508.06749*, 2015.

[16] Anuj Srivastava and Eric P Klassen, *Functional and shape data analysis*, Springer, 2016.

[17] Serge Lang, *Fundamentals of differential geometry*, vol. 191, Springer Science & Business Media, 2012.