A Two-Step Geometric Framework For Density Modeling

Sutanoy Dasgupta, Debdeep Pati and Anuj Srivastava

Department of Statistics, Florida State University

Supplementary Material

We include the theoretical results and derivations of the convergence rates in the sections S1 and S2. Section S3 contains some implementation techniques. We include some pictorial details on the simulation study performed on unconditional density estimation(in §4 of the manuscript) in Section S4. Section S5 contains some additional simulation study to study the properties of the density estimator. Section 6 discusses some properties of the framework and the geometric approach.

S1 Theoretical Results

In this Section we present a detailed discussion on theoretical results and derive properties which lead to the asymptotic convergence rate of the proposed estimator as presented in §3 of the manuscript. First, we introduce some notations and definitions. Recall that $\mathscr{F}$ is the space of all univariate *pdf*s strictly positive on $[0, 1]$ and zero elsewhere. We have represented an arbitrary *pdf* as a function of the coefficients *w.r.t* a basis set of the tangent space. Since we are performing maximum likelihood estimation over an approximating space $\mathscr{F}_n$ for a sample of size n, our estimator is a sieve MLE, discussed in Wong & Shen (1995) and defined in §3 of the manuscript.

To control the approximation error, Wong & Shen (1995) introduces a family of discrepancies. They define $\delta_n(f_0, \mathscr{F}_n) = \inf_{f \in \mathscr{F}_n} \rho(f_0, f)$, called the ρ-approximation error at f_0. The control of the approximation error of $\mathscr{F}_n$ at f_0 is necessary for obtaining results on the convergence rate for sieve MLEs. We follow Wong & Shen (1995) to introduce a family of indexes of discrepency in order to formulate the condition on the approximation error of $\mathscr{F}_n$. Let

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

For two strictly positive densities p and f, set $x = p/f$ and define $\rho_\alpha(p, f) = E_p g_\alpha(X) = \int p g_\alpha(p/f)$. We define $\delta_n(\alpha) = \inf_{f \in \mathscr{F}_n} \rho_\alpha(f_0, f)$. We use $\alpha = 1$ for our results. Then $\delta_n(1) = \inf_{f \in \mathscr{F}_n} \int (f_0 - f)^2/f$, the Pearson's χ^2 number.

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

Let $\|\cdot\|_r$ denote $\mathbb{L}^r$ norm between functions and $\|x\|$ denotes $\sqrt{\int_0^1 x^2(t)dt}$ for functions. For two densities f_1 and f_2, we define the Hellinger metric between f_1 and f_2 as $H(f_1, f_2) = \|f_1^{1/2} - f_2^{1/2}\|_2$. We call a finite set $\{(f_j^L, f_j^U), j = 1, \ldots, N\}$ a

Hellinger u-bracketing of $\mathscr{F}_n$ if $\|f_j^{L^{1/2}} - f_j^{U^{1/2}}\|_2 \leq u$ for $j = 1, \ldots, N$, and for any $p \in \mathscr{F}_n$, there is a j such that $f_j^L \leq p \leq f_j^U$. Let $H(u, \mathscr{F}_n)$ be the Hellinger metric entropy of $\mathscr{F}_n$, defined as the cardinality of the u-bracketing of $\mathscr{F}_n$ of the smallest size. Throughout, c_1 and c_2 have been used to represent coefficient vectors in the tangent space of the Hilbert sphere for some fixed basis set corresponding to warping function that acts on f_p. c_0 denotes a coefficient vector corresponding to the true density denoted by $f_0 \in \mathscr{F}$. l_1, l_2, l_3 and l_4 are used to indicate specific constants. Also, $M_1, M_2, M_3, \ldots,$ have been used to represent generic constants whose value can change from step to step but is independent of other terms in the expressions.

Let f_1 and f_2 be two *pdfs* on $\mathscr{F}_n$ with corresponding cumulative distribution functions F_1 and F_2. Let f_p be the initial density estimate on $\mathscr{F}_p$ such that f_p is strictly positive and Lipschitz continuous with cumulative distribution function F_p. Let $\gamma_1 = {F_p}^{-1} \circ F_1$ and $\gamma_2 = {F_p}^{-1} \circ F_2$. Let $c_1 = (c_{11}, \ldots, c_{1k_n})^{\mathrm{T}}$ and $c_2 = (c_{21}, \ldots, c_{2k_n})^{\mathrm{T}}$ be coefficients associated with two elements of $T_{\mathbf{1}}(\mathbb{S}_\infty)$ corresponding to the tangent space representation of γ_1 and γ_2, that is, $c_1 \in \mathcal{C}_{\gamma_1}$ and $c_2 \in \mathcal{C}_{\gamma_2}$. Here $\mathscr{F}_n$, $\mathcal{C}_\gamma$ and k_n are as introduced in §2 of the manuscript. Then the following Lemma bounds the norm difference of f_1 and f_2 with the norm difference in the coefficients.

Proposition 1. $|f_1 - f_2| \leq M_0 \sum_{i=1}^{k_n} |c_{1i} - c_{2i}|$ *where* $M_0 > 0$ *is a constant.*

Proof. c_1 and c_2 are the coefficients associated with two elements v_1 and v_2 of $T_{\mathbf{1}}(\mathbb{S}_\infty)$, defined in §2 of the manuscript and let q_1 and q_2 represent the corresponding elements on the Hilbert sphere. Then there exists $M_1 \in \mathbb{R}$ such that $|B_i| < M_1$, where B_i is the ith basis function, $i = 1, 2, \cdots, k_n$. Let $v_1 = \sum_{i=1}^{k_n} c_{1i}B_i$, $v_2 = \sum_{i=1}^{k_n} c_{2i}B_i$. Then $v_1, v_2 \in T_{\mathbf{1}}(\mathbb{S}_\infty)$ with $\|v_1\| < \pi$ and $\|v_2\| < \pi$. Hence we have

$$(v_1 - v_2)(t) = \sum_{i=1}^{k_n}(c_{1i} - c_{2i})B_i(t) < M_1 \sum_{i=1}^{k_n} |c_{1i} - c_{2i}|$$

$$\|v_1 - v_2\| = \sqrt{\int_0^1 (v_1 - v_2)^T(v_1 - v_2)dt} < M_3\sqrt{\sum_{i=1}^{k_n}(c_{1i} - c_{2i})^2} < M_1 \sum_{i=1}^{k_n} |c_{1i} - c_{2i}|$$

Next since $x \mapsto \|x\| = \sqrt{\int_0^1 x^2(t)dt}$ and $x \mapsto \cos(x)$ are Lipschitz continuous, we have

$$|\cos\|v_1\| - \cos\|v_2\|| < M_2|\|v_1\| - \|v_2\|| < M_1 \sum_{i=1}^{k_n} |c_{1i} - c_{2i}| \tag{S1.1}$$

Next note that $x \mapsto \sin(x)/x$ is Lipschitz continuous. Hence we have

$$\left\|\frac{\sin\|v_1\|}{\|v_1\|} - \frac{\sin\|v_2\|}{\|v_2\|}\right\| < M_2|\|v_1\| - \|v_2\|| < M_1 \sum_{i=1}^{k_n} |c_{1i} - c_{2i}| \tag{S1.2}$$

Noting that

$$|q_1(t) - q_2(t)| < |\cos\|v_1\| - \cos\|v_2\|| + \left|\frac{\sin\|v_1\|}{\|v_1\|}v_1(t) - \frac{\sin\|v_2\|}{\|v_2\|}v_2(t)\right|$$

we have, combining equations S1.1 and S1.2,

$$\|q_1 - q_2\|_1 < M_1 \sum_{i=1}^{k_n} |c_{1i} - c_{2i}| \tag{S1.3}$$

Now consider $Q = q^2$. Observe that

$$\begin{aligned}(Q_1 - Q_2)(t) &= {q_1}^2(t) - {q_2}^2(t) = (q_1(t) - q_2(t))(q_1(t) + q_2(t)) \\ &= (\cos|v_1| + \cos|v_2| + \frac{\sin|v_1|}{|v_1|}v_1(t) + \frac{\sin|v_2|}{|v_2|}v_2(t))(q_1(t) - q_2(t)).\end{aligned}$$

Now $(\cos\|v_1\| + \cos\|v_2\| + \frac{\sin\|v_1\|}{\|v_1\|}v_1(t) + \frac{\sin\|v_2\|}{\|v_2\|}v_2(t))$ is a bounded function. Hence $\|Q_1 - Q_2\|_1 < M_1\sum_{i=1}^{k_n}|c_{1i} - c_{2i}|$ using equation S1.3. Now we have $\gamma_i(t) = \int_0^t Q_i(u)du, t \in [0,1], i = 1,2$. Then

$$|\gamma_1(t) - \gamma_2(t)| = \left|\int_0^t \Big(Q_1(u) - Q_2(u)\Big)du\right| < \int_0^t |Q_1(u) - Q_2(u)|du \leq \|Q_1 - Q_2\|_1$$

Since f_p is Lipschitz continuous and strictly positive density on $[0,1]$, we have

$$\|f_p(\gamma_1) - f_p(\gamma_2)\|_1 < M_4\|\gamma_1 - \gamma_2\|_1$$

Consider $|f_1 - f_2| = |f_p(\gamma_1).\dot{\gamma}_1 - f_p(\gamma_2).\dot{\gamma}_2|$. Keeping in mind that $Q = \dot{\gamma}$, we have

$$\begin{aligned}
|f_1(t) - f_2(t)| &= |f_p(\gamma_1(t)).Q_1(t) - f_p(\gamma_2(t)).Q_2(t)| \\
&= |f_p(\gamma_1(t)).Q_1(t) - f_p(\gamma_2(t)).Q_1(t) + f_p(\gamma_2(t)).Q_1(t) - f_p(\gamma_2(t)).Q_2(t)| \\
&\leq |Q_1(t)|M_1\|\gamma_1 - \gamma_2\|_1 + |f_p(\gamma_2(t))|\|Q_1 - Q_2\|_1 \\
&\leq M_2\|\gamma_1 - \gamma_2\|_1 + M_3\|\gamma_1 - \gamma_2\|_1 < M_0\sum_{i=1}^{k_n}|c_{1i} - c_{2i}|.
\end{aligned}$$

Therefore we have $|f_1 - f_2| < M_0\sum_{i=1}^{k_n}|c_{1i} - c_{2i}|$ for some fixed $M_0 > 0$. □

Remark 1. *It follows that $H(f_1, f_2) < M_1\sqrt{\|f_1 - f_2\|_1} < M_1\sqrt{\sum_{i=1}^{k_n}|c_{1i} - c_{2i}|} < l_1\sqrt{\max_{1\leq i\leq k_n}|c_{1i} - c_{2i}|}$ for some fixed $l_1 > 0$ where $H(f_1, f_2)$ is the Hellinger metric between two densities f_1 and f_2.*

The following Lemma provides a bound for the Hellinger metric entropy for $\mathscr{F}_n$.

Lemma 1. *1. There exists positive constants C_3 and C_4 and a positive $\epsilon < 1$ such* *that,*

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} H^{1/2}(\frac{u}{C_3}, \mathscr{F}_n)du \leq C_4 n^{1/2}\epsilon^2, \tag{S1.4}$$

2. *There exists positive constants* C_1 *and* C_2 *such that for any* $\epsilon > 0$*,*

$$P^*\left(\sup_{\{\|p^{1/2}-p_0^{1/2}\|_2 \geq \epsilon, p\in\mathscr{F}_n\}} \prod_{i=1}^n p(Y_i)/f_0(Y_i) \geq exp(-C_1 n\epsilon^2)\right) \leq 4\, exp(-C_2 n\epsilon^2)$$

Proof. We note that $H(f_1, f_2) \leq l_1\sqrt{\max_{1\leq i\leq k_n}|c_{1i} - c_{2i}|}$ for some $l_1 > 0$ following Remark 1. So finding a δ covering for $\mathscr{F}_n$ is equivalent to finding an $l_1\sqrt{\delta}$ covering for the space of coefficients in the tangent space using L_∞ norm. Let us have a closer look at the space of coefficients. We have $\|v\| < \pi$ for tangent space representation of Γ, which is equivalent to $\|c\|_2 \leq l_3$,say. Therefore $\mathscr{F}_n \equiv \{c \in \mathbb{R}^{k_n} : \|c\|_2 \leq l_3\} = \mathcal{C}^{k_n}$,say. Then $\mathcal{C}^{k_n} \subset \{c \in \mathbb{R}^{k_n} : \|c\|_\infty \leq l_4\} \equiv \{c \in \mathbb{R}^{k_n} : |c_i| \leq l_4 \forall i = 1, \ldots, k_n\} = \mathscr{C}_0$, say. Now $\mathscr{C}_0$ is a compact set with $\mathcal{C}^{k_n}$ as a compact subset. Therefore the covering number N for $\mathcal{C}^{k_n}$ would be less than the covering number for $\mathscr{C}_0$. Since $\mathscr{C}_0 \equiv \{[-l_4, l_4]^{k_n}\}$, we have the covering number for $\mathscr{C}_0$ as $(\frac{2l_4}{l_1\sqrt{\delta}})^{k_n}$. We obtain this by partitioning the interval $[-l_4, l_4]$ into pieces of length $l_1\sqrt{\delta}$ for each coordinate so that the partition of $\mathscr{C}_0$ is reached through cross product. Then in each equivalent class of the partition of $\mathscr{C}_0$ we will have $\|c_1 - c_2\|_\infty \leq l_1\sqrt{\delta}$ which is equivalent to $H(f_1, f_2) \leq \delta$. So we have the metric entropy for $\mathscr{F}_n = H(., \mathscr{F}_n) = H(u, \mathscr{F}_n) < k_n \log l/u$, where $l = 2l_4$ and $u = l_1\sqrt{\delta}$.

Now,

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} H^{1/2}(\frac{u}{l_3}, \mathscr{F}_n)du \leq \sqrt{k_n}\int\sqrt{\log(l_0/u)}du \leq \sqrt{k_n\log(M/\epsilon^2)}(\sqrt{2}\epsilon - \epsilon^2/256)$$

where $l_0 = l_3 l$ and $M = 2^8 l_0$. For the existence of an ϵ_n that satisfies (S1.4) we need an ϵ_n less than 1 that satisfies

$$\sqrt{k_n\log(M/\epsilon^2)}(\sqrt{2}\epsilon - \epsilon^2/256) \leq C_4 n^{1/2}\epsilon^2 \tag{S1.5}$$

But this inequality holds at $1-$ and hence there exists a smallest $\epsilon_n < 1$ that satisfies S1.5. The Part 2 of the lemma follows directly from Theorem 1 in Wong & Shen (1995). □

Lemma 2. *Under the above notations and assumptions, there exists a positive constant C_5 such that $\delta_n(1) = C_5 n^{-2\beta/(2\beta+1)}$.*

Proof. $\delta_n(1) = \inf_{f\in\mathscr{F}_n}\rho_1(f_0, f) = \inf_{f\in\mathscr{F}_n}\int p_0 g_1(f_0/f). = \inf_{f\in\mathscr{F}_n}\int\frac{(f_0-f)^2}{f}$ be the Pearson's χ^2 number. Now, f_0 is assumed to belong a Hölder space of order $\beta > 0$ (Assumption 2). By Proposition 1, it is straightforward to show that if $k_n = l_1 n^{1/(2\beta+1)}$ then $\inf_{f\in\mathcal{F}_n}\|f_0 - f\|_\infty \leq l_2 n^{-\beta/(2\beta+1)}$ for some arbitrary constants l_1 and l_2 . This 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). Also, f_0 is assumed to be strictly positive. That is, there exists a positive number d such that $\inf\limits_{t\in[0,1]} p(t) \geq 2d$. Let $\hat{f} = \underset{f\in\mathscr{F}_n}{\operatorname{arginf}}\|f_0 - f\|_\infty$. Then it follows that for a large enough n, $\inf\limits_{t\in[0,1]} \hat{f}(t) \geq d$. Then we have, $\delta_n(1) = \inf_{f\in\mathscr{F}_n} \int \frac{(f_0-f)^2}{f} dt < \inf_{f\in\mathscr{F}_n} \|f_0 - f\|_\infty^2/d = C_5 n^{\frac{-2\beta}{2\beta+1}}$ for some $C_5 > 0$. □

S1.1 Proof of Theorem 1

We have from equation S1.5 $\sqrt{k_n \log(M/\epsilon^2)}(\sqrt{2}\epsilon - \epsilon^2/256) < \sqrt{k_n \log(M/\epsilon^2)}\sqrt{2}\epsilon$. So for an upper bound of the smallest root we can solve the equation $\sqrt{k_n \log(M/\epsilon^2)}\sqrt{2}\epsilon = C_4 n^{1/2}\epsilon^2$. Let ϵ_n be of the form $\sqrt{M}n^{-\gamma}(\log n)^t, \gamma > 0$, and, let $k_n = n^\Delta$, $\Delta < 1$ Then $\log\left(M/{\epsilon_n}^2\right) = 2\gamma \log n - 2t \log\log n \leq 2\gamma \log n$.

So for an upper bound of the smallest root we can solve the equation

$\sqrt{k_n 2\gamma \log n}\sqrt{2}\epsilon = C_4 n^{1/2}\epsilon^2$.

Therefore equating, $n^{\Delta/2}\sqrt{2\gamma}\sqrt{\log n}\sqrt{M}n^{-\gamma}(\log n)^t$ with $C_4 M n^{1/2} n^{-2\gamma}(\log n)^{2t}$, we get $\gamma = \frac{1}{2}(1-\Delta)$, and $t = 1/2$. Thus we have $\epsilon_n = \sqrt{M}n^{\frac{-(1-\Delta)}{2}}\sqrt{\log n}$. We take Δ to be $\frac{1}{2\beta+1}$ to use the theoretical properties of Hölder space of order $\beta > 0$. Therefore $\epsilon_n = \sqrt{M}n^{\frac{-\beta}{2\beta+1}}\sqrt{\log n}$ is an upper bound for the smallest value that satisfies the condition for Lemma 1. Therefore, using the definition given in Theorem 4 in Wong & Shen (1995), we get

$$\epsilon_n^* = \begin{cases} Mn^{-\beta/(2\beta+1)}\sqrt{\log n}, \text{ if } \delta_n(1) < \frac{1}{4}C_1 M^2 n^{-2\beta/(2\beta+1)}\log n, \\ (4\delta_n(1)/C_1)^{1/2}, \text{ otherwise.} \end{cases}$$

But $\delta_n(1) = C_5 n^{-2\beta/(2\beta+1)} < \frac{1}{4}C_1 M^2 n^{-2\beta/(2\beta+1)} \log n$ for $n > \exp(4C_5/M^2C_1)$. Thus for large enough n, $\epsilon_n^* = Mn^{-\beta/(2\beta+1)}\sqrt{\log n}$ and following Theorem 4 of

Wong & Shen (1995) we get

$$P(\|q^{1/2} - p_0^{1/2}\|_2 \geq \epsilon_n^*) \leq 5\text{exp}\big(- C_2 n(\epsilon_n^*)^2\big) + \text{exp}\big(- \frac{1}{4} n C_1 (\epsilon_n^*)^2\big). \qquad \text{(S1.6)}$$

S2 Theoretical Results for Conditional Densities

Here we prove the asymptotic consistency properties of the conditional density estimate. First, we introduce some new notations and borrow some notations from previous discussions and redefine some in the context of conditional density estimation.

Let f_X^m denote the marginal density of X. Let $f_x(y) \equiv f_x(y|X = x)$ be the true conditional density estimator of y given $X = x$. Let $\tilde{f}_x(y|X = x)$ be any conditional density estimate of $f_x(y|X = x)$. Let $\mathcal{X}$ denote the compact support of the predictor variable X and let $\mathcal{Y}$ denote the support of the response variable Y. For simplicity, we assume $\mathcal{Y} = [0, 1]$. Let k_n be the number of basis elements used for the tangent space representation of the warping functions $\gamma \in \Gamma$ for sample size n. Let $f_p(y|X = x)$ be the initial guess of the conditional density of y given x, as discussed in Section 5 of the paper. Define $\mathscr{F}_x$ as the set $\{f_p(\gamma(y)|x)\dot{\gamma}(y), \gamma \in \Gamma\}$. Define the approximating space of densities $\mathscr{F}_{x,n}$ as $\mathscr{F}_{x,n} = \{f_p(\gamma(y)|x)\dot{\gamma}(y)f_X(x)|\gamma \in \Gamma^{k_n}\}$, where Γ^{k_n} is defined in Section 2.1 of the paper. Let w_{i,n,x_0} refer to the weights centered at $x_0 \in \mathcal{X}$ associated with the likelihood function, defined in §5 of the main paper, after normalization. W_{i,n,x_0} refers to the weights before normalization. For instance, the weights were chosen to be Gaussian kernels in the simulated examples. c refers to a generic positive constant that changes value from step to step but do not affect other terms. $l(\gamma|y_i, x_i) = \log f_p(\gamma(y_i)|x_i)\dot{\gamma}(y_i)$ refers to the log likelihood of γ under the assumption $y_i \sim f_{x_i}$. On the other hand, $l(\gamma|y_i, x_0) = \log f_p(\gamma(y_i)|x_0)\dot{\gamma}(y_i)$ refers to the log likelihood of γ given that $y_i \sim f_{x_0}$. For a given $\gamma \in \Gamma$ and $x_0 \in \mathcal{X}$, let $\tilde{L}_{n,x_0}(\gamma) = \sum_{i=1}^{n} \log f_p(\gamma(y_i)|x_i)\dot{\gamma}(y_i)w_{i,n,x_0}$. We drop the x_0 suffix henceforth for easier notation. Let $\hat{\gamma}_{n,x_0} = \operatorname{argmax} \tilde{L}_{n,x_0}(\gamma)$, and $\hat{f}_{x_0,n} = f_p(\hat{\gamma}_n(\cdot)|X = x)\dot{\hat{\gamma}}_n \in \mathscr{F}_{x_0,n}$ be the sieve maximum likelihood density estimate. Let F_x and $F_p(\cdot|X = x)$ be the cumulative distribution functions corresponding to f_x and $f_p(\cdot|X = x)$, respectively. Then $\gamma_{0x} = F_p^{-1}(F_x)$ is the true warping function. For any γ, let v_γ represent the tangent space representation of γ. Also, let $v_{\gamma_{id}} \equiv 0$ be the tangent space representation corresponding to $\gamma_{id}(t) = t$, the identity warping function.

Now, we list some assumptions under which we can prove the consistency of the conditional density estimator and derive an upper bound of the convergence rate:

A1. The marginal density of X, denoted by f_X^m is Lipschitz continuous, and is bounded away from zero on its support.

A2. The support $\mathcal{Y}$ of the response variable Y is $[0, 1]$.

A3. The true conditional density $f_{x_0} : [0, 1] \to \mathbb{R}^+$ is continuous and strictly positive.

A4. The initial shape $f_p : [0, 1] \to \mathbb{R}^+$ is bounded away from zero, and is Lipschitz continuous.

A5. For any target location $x_0 \in \mathcal{X}$, the weights $w_{i,n}$ satisfy the following properties:

a. There exists a positive sequence $h_n \to 0$ with $nh_n \to \infty$ such that for all $x_0, x \in \mathcal{X}, y \in \mathcal{Y}$, and $n \geq 1$, we have $\sup_{\gamma \in \Gamma^{k_n}, |x - x_0| < h_n} ||[l(\gamma|y, x_0) - l(\gamma|y, x)]|| < ch_n$.

b. $w_{i,n} \leq n^{-1}o(1)$ for all $\{x_i : |x_i - x_0| > h_n\}$.

c. The number of observations in the h_n-neighborhood of x_0 is of the order of nh_n, for all $x_0 \in \mathcal{X}$, where the h_n-neighborhood is $[x_0 - h_n, x_0 + h_n]$.

d. $1 \leq \{\max_{i:|x_i - x_0| < h_n}\{w_{i,n}\}\{\min_{i:|x_i - x_0| < h_n}\{w_{i,n}\}^{-1} \leq C_{0n}$, where C_{0n} is a bounded positive sequence with $C_{0n} \to 1$ as $n \to \infty$.

A6. The true conditional density $f_{x_0}(\cdot|X = x_0)$ either belongs to Hölder or Sobolev space of order β.

Assumptions **A1** and **A5.c** make sure that for any $x_0 \in \mathcal{X}$, the number of observations in the neighborhood h_n of x_0 grows at the same rate (upto constants) as the sample size increases. The Assumption **A2** is for simplicity of analysis. Assumption **A4** ensures that the initial guess is in the correct space of densities. Some simple examples of initial guesses can be truncated gaussian densities, any standard conditional density estimate, and so on. Assumption **A5.a** is a very important assumption which states that the true conditional density of the observations in a neighborhood around the target location is "not too far away" from that of the target location. Further, this discrepancy of the observations in the neighborhood decreases to zero as the sample size increases. Without this assumption, it is not possible to derive a good estimate of the true density. Assumption **A5.b** and **A5.d** ensures that the weights associated with the observations in the neighborhood of the target location are asymptotically same, whereas the weights outside the neighborhood collapse to zero at a faster rate. This ensures that the weighted likelihood function asymptotically behaves like an unweighted likelihood function of observations in a neighborhood around x_0.

Theorem 1. *Under assumptions A1-A5, for any fixed* $\epsilon > 0$ *and* $x_0 \in \mathcal{X}$, $P(\|f_{x_0}^{1/2} - \hat{f}_{x_0}^{1/2}\|_2 \geq \epsilon) \to 0$.

Proof. Note that we have $\sum_{i=1}^{n} w_{i,n} = 1$. Let $s_{x_0}(h_n) \sim nh_n$ denote the number of observations which lie in the h_n neighborhood of x_0. Let $w_{min} = \min\limits_{i:|x_i - x_0| < h_n} \{w_{i,n}\}$. Then we have, $(1 - o(1))/C_{0n} s_{x_0}(h_n) \leq w_{min} \leq (1 - o(1))/s_{x_0}(h_n)$. Thus, we have by sandwich theorem, $w_{i,n} \to 1/s_{x_0}(h_n)$ as $n \to \infty$ for $\{i : |x_i - x_0| < h_n\}$.

Thus, $\tilde{L}(\gamma) = \sum_{i=1}^{s_{x_0}(h_n)} w_{i,n} l(\gamma|y_i, x_0) - \eta_n$, where $\eta_n = \sum_{i=1}^{s_{x_0}(h_n)} w_{i,n} |l(\gamma|y_i, x_i) - l(\gamma|y_i, x_0)| + \sum\limits_{i:|x_i - x_0| > h_n} |w_{i,n} l(\gamma|y_i, x_i)| <= C_{0n} \sum_{i=1}^{s_{x_0}(h_n)} (1/s_{x_0}(h_n)) c h_n + o(1) <= c h_n + o(1) \to 0$. This follows from the fact that $|l(\gamma)|y_i, x_i)| < |l(\gamma|y_i, x_i) - l(\gamma_{id}|y_i, x_i)| + |l(\gamma_{id}|y_i, x_i) < c\|v_\gamma - v_{\gamma_{id}}\| + |f_p(y_i|x_i)|$ following the steps of Proposition 1 in Section S1. Thus $|l(\gamma|y_i, x_i)|$ is uniformly bounded and $\lim\limits_{n\to\infty} \underset{\gamma \in \Gamma^{k_n}}{\text{argmax}} \tilde{L}_{n,x_0}(\gamma) = \lim\limits_{n\to\infty} \underset{\gamma \in \Gamma^{k_n}}{\text{argmax}} \sum_{i=1}^{s_{x_0}(h_n)} 1/(s_{x_0}(h_n)) l(\gamma|y_i, x_0)$. Thus, we obtain $P(\|f_{x_0}^{1/2} - \hat{f}_{x_0}^{1/2}\|_2 \geq \epsilon) \to 0$ using the standard consistency properties of sieve MLEs that follows from the analysis in Wong & Shen (1995). □

Corollary 1. *Let* $\epsilon_n^* = c s_{x_0}(h_n)^{-\beta/(2\beta+1)} \log s_{x_0}(h_n)$. *Under assumptions A1-A6,* *there exists constants C_1 and C_2 such that if $\eta_n \leq c\epsilon_n^{*2}/2$,*

$P(\|f_{x_0}^{1/2} - \hat{f}_{x_0}^{1/2}\|_2 \geq \epsilon) \leq 5\exp(-C_2 s_{x_0}(h_n)\epsilon_n^{*2}) + \exp(-s_{x_0}(h_n)C_1\epsilon_n^{*2}).$

This result follows directly using the theory for unconditional density estimation using $s_{x_0}(h_n)$ observations. Note that the rate of convergence is slower than minimax rate. It depends on the number of significant observations $s_{x_0}(h_n)$ in the neighborhood of x_0, and the dicrepancy of observations η_n with respect to the target location.

Corollary 2. *Suppose $\gamma_{0,x_0} \in \Gamma^{k_0}$ where k_0 is a known finite number. Let the normalized weights satisfy that $w_{i,n} = 0$ for all $\{x_i : |x_i - x_0| > h_n\}$, and uniform in the h_n neighborhood. Then, if $\epsilon_n = cn^{-1/3}$ for some constant c, $P(\|f_{x_0}^{1/2} - \hat{f}_{x_0}^{1/2}\|_2 \geq \epsilon_n) \leq$* $5\exp(-n\epsilon_n^2)$.

Proof. Note that $\eta_n = \sum_{i=1}^{s_{x_0}(h_n)} w_{i,n}|l(\gamma|y_i, x_i) - l(\gamma|y_i, x_0)| \sim h_n$. Using $s_{x_0}(h_n)$ observations for Theorem 2 in Wong & Shen (1995), we get $\epsilon_n = s_{x_0}(h_n)^{-1/2}$ as the convergence rate. This is the fairly standard $n^{-1/2}$ convergence rate for MLEs in finite dimensions. Choosing $h_n = n^{-1/3}$, we get $\epsilon_n = cn^{-1/3}$. The statement then follows from directly applying Theorem 2 in Wong & Shen (1995). Interestingly, this convergence rate agrees with Hu (1997) who obtained $(\sum w_{i,n}^2)^{1/2}$ as the convergence rate for relevance weighted MLEs. □

S3 Estimation Algorithm

In this section we outline the estimation procedure and discuss some of the implementation issues. We discretize density functions using a dense uniform partition, $T = 100$ equidistant points over the interval $[0, 1]$. For approximating derivatives of a function, for example $\dot{\gamma}$ for a warping function γ, we use the first-order differences. The integrals are approximated using the trapezoidal method.

For optimizing log-likelihood function according to Equation 2.5 of the manuscript, we use the function *fminsearch* in `MATLAB` for our experiments. The *fminsearch* function uses a very efficient grid search technique to find the optimal values of coefficients $\{c_j\}$, corresponding to the chosen basis elements, to approximate the optimal warping function γ. However, *fminsearch* function can get stuck in locally-optimal solutions in some situations. To alleviate this problem we use an iterative, multi-resolution approach as follows. We start the optimization using a small number of basis elements J with $c = \mathbf{0}$, the point that maps to $\gamma_{id} \in \Gamma$ under H. This implies a low-resolution search and low-dimensional search space $\mathbb{R}^J$. Then, at each successive iteration we increase the resolution by increasing J and use the previous solution as the initial condition (with the additional components set to zero) for the next stage. This slow increase in J, while continually improving the optimal point c, performs much better in practice than using a large value of J directly in *fminsearch*.

Another important numerical issue is the final choice of J. For a fixed sample of size n, a large value of J may lead to overfitting and $\hat{f}$ being a rough function. Also, a large value of J makes it harder for the search procedure to converge to an optimal solution. Efromovich (2010) and the references there in discusses different data-driven methods to choose the number of basis elements, by considering the number of basis elements itself as a parameter. We take a different data-driven approach for selecting the desired number of basis elements. Using a predetermined maximum number of basis points, we navigate through increasing number of basis elements and at each step, we compute the value of the Akaike's Information Criterion (AIC) and choose the number of basis elements that results in the best value of the AIC, penalizing the number of basis functions used. We summarize the full procedure in **Algorithm 1**.

Algorithm 1 Improving solutions using *fminsearch* by tweaking the starting points

```
i. Start with a low number of basis elements, say $J$
ii. Use 0 vector as the starting point and find the solution d using fminsearch.
iii. Increase the number of basis elements, say $J_1$ more basis elements.
iv. Use [0,0] and [d,0] as two starting points. Compare the AIC for the two cases and choose the solution with better AIC value. Call the solution d the optimal solution.
v. If the number of basis elements exceeds a predetermined large number, stop. Else go to step iii.
```

Experimental results show that Bayesian Information Criterion (BIC) overpenalizes the number of basis elements used and, therefore, some sharper features of the true density are lost in the estimate. So the experiments presented in the following sections use only the AIC penalty.

S4 Simulation Studies I

Next, we elaborate on the results from experiments on univariate unconditional density estimation procedure involving two simulated datasets, from Section 5 in the manuscript. The computations described here are performed on an Intel(R) Core(TM) i7-3610QM CPU processor laptop, and the computational times are reported for each experiment. We compare the proposed solution with two standard techniques: (1) kernel density estimates with bandwidth selected by unbiased cross validation method, henceforth referred to as *kernel(ucv)*, (2) a standard Bayesian technique using the function *DPdensity* in the R package `DPPackage`. The Bayesian approach naturally has a longer run-time. For both the simulated examples, we use 2000 MCMC runs with 500 iterations as burn in period for the Bayesian technique. We compare the methods both in terms of numerical performance and computational cost. Here we illustrate the performance of the various methods using a representative simulation. We highlight the performance improvement over an (misspecified) initial parametric and nonparametric density estimate brought about by warping. For the initial parametric estimate we have chosen a normal density truncated to $[0, 1]$ with mean and standard deviation estimated from the sample. For the initial nonparametric estimate, we used inbuilt `MATLAB` function *ksdensity*.

S4.1 Example 1

We borrow the first example from Tokdar (2007) and Lenk (1991), where $f_0 \propto 0.75\text{exp}(\text{rate} = 3) + 0.25\mathcal{N}(0.75, 2^2)$, a mixture of exponential and normal density truncated to the interval $[0, 1]$: We generate $n = 100$ observations to study estimation performance. Here we use Meyer wavelets as the basis set for the tangent space representation of γs. We use Algorithm 1 to adaptively choose the number of basis elements. For this example, we start with 8 basis elements and introduce 7 basis elements at each step upto a maximum of 29 basis elements, akin to multiresolution analysis. Also, we use an unpenalized log likelihood for optimization.

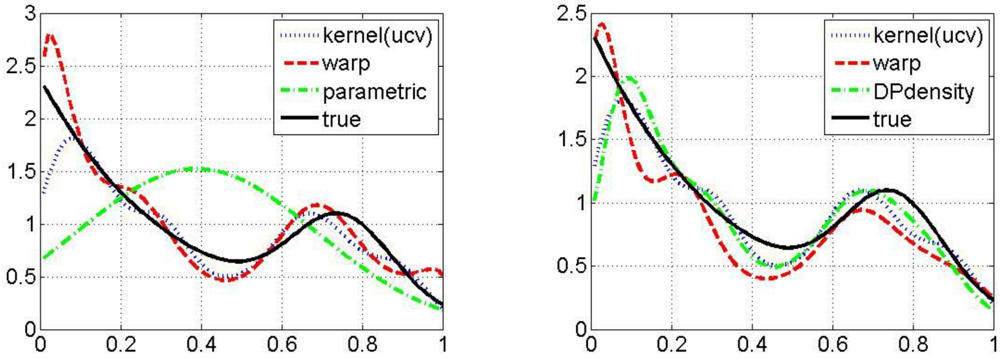

Figure 1: *The left panel compares the warped estimate $\hat{f}$ with other estimates when f_p is parametric. The middle panel shows the corresponding evolution of the negative of log-likelihood function during optimization. The right figure compares the warped estimate with others when f_p is ksdensity.*

Figure 1 (left panel) shows a substantial improvement in the final warped estimate over the initial parametric estimate. Incidentally, it also does a better job in capturing the left peak as compared to the *kernel(ucv)* method. Standard kernel methods need additional boundary correction techniques to be able to capture the density at the boundaries, as studied in Karunamuni & Zhang (2008) and the references therein. However the warped density seems to perform better estimation near the boundaries compared to the other techniques. The right panel displays the warped result when using *ksdensity* output as the initial estimate. It also provides solutions obtained using *kernel(ucv)* and *DPdensity*. Once again, this warped estimate provides a substantial improvement over the initial solution.

S4.2 Example 2

For the second example we take Example 10 from Marron & Wand (1992), which uses a claw density: $f_0 = \frac{1}{2}\mathcal{N}(0,1) + \sum_{l=0}^{4} \frac{1}{10}\mathcal{N}(\frac{l}{2} - 1, (0.1)^2)$. We estimate the

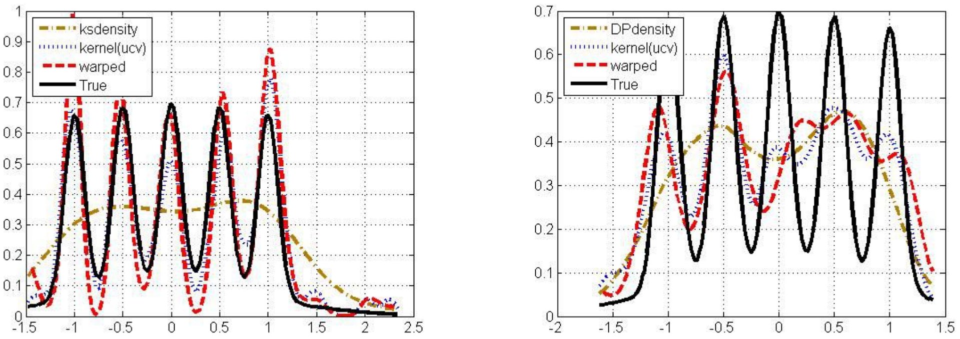

Figure 2: *The left panel shows the improvement over initial ksdensity estimate. Both kernel(ucv) and warped estimate have a good performance here. The right panel shows that all the methods fail to capture all the peaks. Kernel(ucv) performance is very similar to the warped estimate.*

domain boundaries and unlike the previous example, instead of fixing the number of tangent basis elements, we employ Algorithm 1 described in Section S3 to find the optimal number of basis elements based on the AIC, with a maximum allowed value of 40 basis elements. We start with 5 basis elements, and add 5 more basis elements at each step. Consequently, the computation cost increases compared to the previous example because we consider more models with different number of parameters. The choice of the initial number of basis elements and number of basis elements to add (step size) is left for the user to decide. Using finer step size tends to improve the practical performance at the cost of higher computational cost.

S5 Simulation Studies II

In this section we perform some additional experiments and comparisons illustrating the properties of the proposed estimator.

S5.1 Effect of the initial shape

In Section 3 of the paper we have presented the asymptotic convergence rate of the estimator to the true density via the convergence of the warping function estimate to the “true” warping function. Naturally, the notion of a “true” warping function requires a fixed initial shape (denoted by f_p), and we have shown that the convergence rate of the warping function estimate to the true warping function is independent of the choice of f_p up to constants.

As a result, the computationally cheapest choice of the initial shape is the uniform density shape. However, the proposed method can also be used to improve the performance of misspecified parametric or nonparametric density estimates. In this section, we consider three different choices of the initial shape f_p, and evaluate the improvement over the initial shape, and compare the performances of the final density estimates given these different initial choices. The three choices of f_p considered are (1) Gaussian, with parameters estimated from the data, (2) Uniform on th unit interval, and (3) a kernel density estimate with bandwidth chosen by Silverman's thumb rule.

To illustrate the improvement over the initial shapes and compare the final performances, we consider the same two examples as discussed in Section 4 of the main paper. For this study we consider sample size 100 throughout and use Algorithm 1 discussed in the Supplementary Section to obtain the number of basis elements for tangent space representation of the warping functions. We start with 5 basis elements and introduce 5 basis elements at each step up to a maximum of 40 basis elements, and choose the model with the best AIC. Table 1 summarizes the performances of the three different choices of starting points for the mixture of exponential and normal example. Table 2 summarizes the same for the claw example. They indicate that even though the performance of the initial estimates are quite poor (albeit computationally very cheap), the warped estimate provides a very significant improvement via the loss functions considered. Note that, even though the loss functions for the initial shapes are quite different, the loss functions of the final estimate are very similar across different initial shapes. Also, for all the chosen initial shapes, the computational costs for obtaining the final estimate are very similar. In general, it is desirable to have the initial shape f_p as close to the true shape as possible so that the "true" warping function is close to the identity function. Since the identity function is the starting point in our search algorithm, the estimation is easier if the true value is close to the searching point. Hence we advocate using the kernel density with a plug-in bandwidth as a good choice of f_p. This is because the estimate is nonparametric and thus robust to different shapes of the true density and is also computationally cheap.

S5.2 Comparison of boundary performance versus kernel densities

Kernel density estimates have very good performance in the interior of the support of the density. However, for densities with compact support, kernel densities tend to have a lot of bias at the boundaries, and typically need boundary correction techniques to have better performance. In this section we compare the performance of the proposed warped estimate with the traditional kernel method focusing on the edges of a compact support. For this purpose, we use a different loss function $\mathbb{L}^B$, defined as follows. Let the support of the density be $[0, 1]$. Let f_0 be the true density and $\hat{f}$ be the density

Table 1: *A comparison of the performances for mixture of exponential and normal example.*

Method:		Initial Shape			Warped Estimate		
Choice of f_p	Norm	Mean	std.dev	Time	Mean	std.dev	Time
Gaussian	$\mathbb{L}^1$	50.37	2.11	< 1 sec	20.93	5.59	5 sec
	$\mathbb{L}^2$	6.36	0.23		2.94	0.67	
	$\mathbb{L}^\infty$	1.75	0.08		1.13	0.28	
Uniform	$\mathbb{L}^1$	34.81	0	< 1 sec	19.33	5.39	5 sec
	$\mathbb{L}^2$	4.64	0		2.64	0.72	
	$\mathbb{L}^\infty$	1.33	0		0.95	0.36	
Kernel	$\mathbb{L}^1$	27.74	3.59	< 1 sec	19.70	5.49	5 sec
	$\mathbb{L}^2$	3.87	0.52		2.77	0.70	
	$\mathbb{L}^\infty$	1.40	0.13		1.05	0.32	

estimate. Then $\mathbb{L}^B(\hat{f}) = (f_0(0) - \hat{f}(0))^2 + (f_0(1) - \hat{f}(1))^2$. For illustration, we use the first example again, $(1) f_0 = [0.75\text{exp}(\text{rate} = 3) + 0.25\mathcal{N}(0.75, 0.25)]I_{[0,1]}$, and $(2) f_0(t) \propto t/5 + (0.5 - t)^2, t \in [0, 1]$, a density with peaks at both the boundaries. Table 3 summarizes the results and shows that the warped estimate has significantly better boundary bias properties than a standard kernel method in all the cases.

S5.3 Choice of basis elements

We have performed the experiments with Fourier basis elements, wavelets and cosine basis elements. The overall performance of the proposed estimator was very similar

Table 2: *A comparison of the performances for claw density example.*

Method:		Initial Shape			Warped Estimate		
Choice of f_p	Norm	Mean	std.dev	Time	Mean	std.dev	Time
Gaussian	$\mathbb{L}^1$	13.44	2.16	< 1 sec	8.57	2.64	22 sec
	$\mathbb{L}^2$	1.80	0.20		1.21	0.36	
	$\mathbb{L}^\infty$	0.45	0.02		0.38	0.13	
Uniform	$\mathbb{L}^1$	19.95	≈ 0	< 1 sec	8.97	2.52	22 sec
	$\mathbb{L}^2$	2.29	≈ 0		1.25	0.33	
	$\mathbb{L}^\infty$	0.44	≈ 0		0.38	0.11	
Kernel	$\mathbb{L}^1$	13.45	2.05	< 1 sec	8.55	2.72	22 sec
	$\mathbb{L}^2$	1.69	0.14		1.22	0.35	
	$\mathbb{L}^\infty$	0.39	0.03		0.39	0.13	

Table 3: *A comparison of the* $\mathbb{L}^B$ *loss function for the warped estimate and the standard kernel method.*

Setup:		Kernel Estimate		Warped Estimate	
Example	n	Mean	std.dev	Mean	std.dev
$f_0 \propto 0.75\exp(3) + 0.25\mathcal{N}(0.75, 0.25)$	25	2.11	0.96	1.60	1.35
	100	2.03	0.61	1.17	0.73
	1000	2.32	0.29	0.82	0.36
$f_0(t) \propto t/5 + (0.5 - t)^2$	25	2.89	0.98	2.39	3.53
	100	1.90	0.81	0.61	0.91
	1000	0.96	0.44	0.27	0.26

for different choices of basis elements. However, 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 can be a good alternative.

S6 Some Properties of the Geometric Framework

S6.1 Advantages Over Direct Approximations

In the previous section, we have used the geometry of Γ to develop a natural, local flattening of Γ. Other, seemingly simpler, choices are also possible but at some loss in estimation performance. For instance, since any γ can also be viewed as a nonnegative function in $\mathbb{L}^2$ with appropriate constraints, it may be tempting to use $\gamma(t) = \sum_{j=1}^{\infty} c_j b_j(t)$, for some orthogonal basis $\mathcal{B} = \{b_j, j = 1, 2, \dots\}$ of $\mathbb{L}^2[0, 1]$ as in Hothorn, Möst & Bühlmann (2015). This seems easier than our approach as it avoids going through a nonlinear transformations. However, the fundamental issue with such an approach is that Γ is a nonlinear manifold and one cannot technically express and estimate elements of Γ directly using linear representations, not even in a small neighborhood. Hothorn, Möst & Bühlmann (2015) uses Bernstein polynomials, with monotonically increasing coefficients, to represent elements of Γ. However, one does not reach the entire set Γ using such a representation. To be specific, it is easy to find a significant subset of Γ whose elements cannot be represented in this system. As a simple example, consider a $\gamma = \sum_{i=0}^{4} c_i B_{i,4}$ with $c_0 = 0$, $c_1 = 0.4$, $c_2 = 0.3$,$c_3 = 0.5$, $c_4 = 1$ (not satisfying the monotonicity constraint). Here, $B_{i,4}$ refer to the Bernstein basis elements of order 4. Even though this γ is a proper diffeomorphism, it cannot be represented in the system used by Hothorn, Möst & Bühlmann (2015).

Another issue in directly approximating element of Γ that both γ and $\dot{\gamma}$ are present in the final estimate and one needs a good approximation of both of these functions. However, a good approximation of γ does not automatically imply a good approximation of $\dot{\gamma}$. In contrast, the reverse holds true as shown next.

Proposition 2. *For any $\gamma \in \Gamma$, let $\dot{\gamma}_{\text{app}}$ be an approximation of $\dot{\gamma}$, and let γ_{app} be the integral of $\dot{\gamma}_{\text{app}}$. For all $x_0 \in (0,1]$ consider intervals I_{x_0} of the form $[0, x_0]$. Then, on all intervals I_{x_0}, $\|\gamma - \gamma_{\text{app}}\|_\infty \leq \|\dot{\gamma} - \dot{\gamma}_{\text{app}}\|_\infty$.*

Proof: For any $t \in I_{x_0}$, the quantity $|\gamma(t) - \gamma_{\text{app}}(t)| = |\int_0^t \dot{\gamma}(s)ds - \int_0^t \dot{\gamma}_{\text{app}}(s)ds| \leq \int_0^t |\dot{\gamma}(s) - \dot{\gamma}_{\text{app}}(s)|ds \leq \|\dot{\gamma} - \dot{\gamma}_{\text{app}}\|_\infty.t \leq \|\dot{\gamma} - \dot{\gamma}_{\text{app}}\|_\infty.x_0 \leq \|\dot{\gamma} - \dot{\gamma}_{\text{app}}\|_\infty$ $\square$

This proposition states that a good approximation of $\dot{\gamma}$ ensures a good approximation of γ, and supports our approach of approximating γ via the inverse exponential transformation of its SRSF to the tangent space $T_{\mathbf{1}}(\mathbb{S}_\infty^+)$. On the other hand, a direct approximation of γ will need many more basis elements to ensure a good approximation of $\dot{\gamma}$.

S6.2 Estimation of Densities with Unknown Support

So far we have restricted to the interval $[0, 1]$ for representing a *pdf*. However, the framework extends naturally to *pdf*s with unknown support. For that, we simply scale the observations to $[0, 1]$ and carry out the original procedure. Let $X_1, X_2, \ldots, X_n \sim f_0$, where X_is are n independent observations from a density f_0 with an unknown support. We transform the data as $Z_i = \frac{X_i - A}{B - A}$, where A and B are the estimated boundaries of the density. Following Turnbull & Ghosh (2014), we take $A = X_{(1)} - s_X/\sqrt{n}$, and $A = X_{(n)} + s_X/\sqrt{n}$, where $X_{(1)}$ and $X_{(n)}$ are the first and last order statistics of X, and s_X is the sample standard deviation of the observed samples. Using the scaled data, we can find the estimated *pdf* f_w on $[0, 1]$ and then undo the scaling to reach the final solution. Turnbull & Ghosh (2014) provide a justification for the choice of A and B as the estimates for the bounds of the density. They also discuss an alternate way of estimating the boundaries using ideas presented in De Carvalho (2011), and suggest that the Carvalho method produces wider and more conservative boundary estimates.

Finally, using the fact that any piecewise continuous density function, with support $\mathbb{R}$ and range $\mathbb{R}_{\geq 0}$, can be approximated to any desired degree by a strictly positive density function on some bounded interval $[A, B]$ (under $\mathbb{L}^2$ norm, for example) , we can extend our method to this larger class of functions.

Bibliography

De Carvalho, M. (2011), 'Confidence intervals for the minimum of a function using extreme value statistics', *International Journal of Mathematical Modelling and Numerical Optimisation* **2**(3), 288–296.

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

Hothorn, T., Möst, L. & Bühlmann, P. (2015), 'Most likely transformations', *arXiv preprint arXiv:1508.06749* .

Hu, F. (1997), 'The asymptotic properties of the maximum-relevance weighted likelihood estimators', *Canadian Journal of Statistics* **25**(1), 45–59.

Karunamuni, R. J. & Zhang, S. (2008), 'Some improvements on a boundary corrected kernel density estimator', *Statistics & Probability Letters* **78**(5), 499–507.

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

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

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

Triebel, H. (2006), ‘Theory of function spaces. iii, volume 100 of monographs in mathematics’, *BirkhauserVerlag, Basel* .

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

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