CAKE:Convex Adaptive Kernel Density Estimation

Abstract

In the main paper we presented results regarding the MSE of CAKE like estimators, and a risk bound for CAKE. In this supplementary material we shall provide complete proofs of both these results.

I. Optimal MSE for "CAKE like" estimators

In this section we consider estimators of the form

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\alpha_{ij}}{h_j^d}k\Big(\frac{x-x_i}{h_j}\Big) \tag{1}$$

where the coefficients $\alpha_{ij} \geq 0$ and satisfy the constraint $\forall i = 1,\ldots,n : \sum_{j=1}^{m}\alpha_{ij} = 1$. Our objective is to calculate the MSE of such an estimator, and the optimal value. We shall call estimators given in equation (1) "CAKE like estimators". We need the following definitions and assumptions.

Definition 1. *Let $\beta, L > 0$. The Hölder class $\Sigma(\beta, L)$ is defined as the set of all functions $f : [0,1]^d \to \mathbb{R}$ which are $l = \lfloor\beta\rfloor$ times differentiable and*

$$|D^l f(x)[\underbrace{h,\ldots h}_{l\ times}] - D^l f(x')[\underbrace{h,\ldots h}_{l\ times}]| \leq L|x-x'|^{\beta-l}|h|^l \ \forall x, x' \in [0,1]^d, h \in \mathbb{R}^d \tag{2}$$

where $\lfloor\beta\rfloor$ is the greatest integer strictly less than β.

Definition 2. *Let $l \geq 1$ be an integer. We say that a kernel $k : \mathbb{R}^d \to \mathbb{R}^d$ has order l if*

$$\int_{u\in\mathbb{R}^d} k(u)\ \mathrm{d}u = 1, \int_{u\in\mathbb{R}^d} u_1^{j_1}u_2^{j_2}\ldots u_d^{j_d}k(u)\ \mathrm{d}u = 0\ \forall j_1,\ldots j_d \geq 0, \sum_{i=1}^{d} j_i \leq l \tag{3}$$

If $d = 1$, then the above condition becomes $\int_{u\in\mathbb{R}} k(u)\ \mathrm{d}u = 1, \int_{u\in\mathbb{R}} u^j k(u)\ \mathrm{d}u = 0\ \forall j = 1\ldots l$.

Assumption 1 (A1). *The set $\mathcal{K}$ has smoothing kernels whose bandwidths $h_j\ \forall j = 1,\ldots,m$ staisfy the constraint $\frac{h_{j_1}}{h_{j_2}} = c_{j_1 j_2}\ \forall j_1, j_2 = 1\ldots m$ where $0 < c_{j_1 j_2} < \infty$ and $h_j \to 0$ as $n \to \infty$ $\forall j = 1\ldots m$.*

Assumption 2 (A2). *The true density function f belongs to the Holder class $\Sigma(\beta, L)$ and the base kernels are of order $l = \lfloor\beta\rfloor$. Also $C_1 \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} k^2(\theta)\ \mathrm{d}\theta < \infty, C_2 \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} |\theta|^\beta k(\theta)\ \mathrm{d}\theta < \infty$.*

Assumption A1 guarantees that as we see more and more samples the bandwidths all tend to 0 at the same rate. Assumption A2 is satisfied for most commonly used smoothing kernels such as a Gaussian kernel or Epanechnikov kernel. The analysis has 3 main steps that can be enumerated as follows.

1) Lemma (1) establishes an upper bound on the bias, variance, and the MSE for CAKE like estimators in terms of α's. The proof techniques used here are fairly standard and similar to ones used in Tsyabkov [1].
2) The next step is to solve an optimization problem P1 (see page 5 of this supplementary material) of minimizing the upper bound on the MSE of CAKE like estimators under convexity constraints on α. This problem doesn't have a closed form solution in general. But we show in Lemmas (4-9) that under assumption A1, A2, and for large enough n it is indeed possible to give a closed form expression for the optimal α. Using these optimal α we calculate the optimal upper bound on MSE in Lemma (10).
3) Finally by spectral analysis in Lemmas (11-2) we are able to investiagte the size of the above derived upper bound on the optimal MSE.
4) The proof of the final result presented in Theorem (3) requires just putting together all the above lemmas.

Fact 1 (Bias-Variance Decomposition). *Let f be the underlying density function. For any estimator $\hat{f}$ let $MSE(x_0) \stackrel{\text{def}}{=} \mathbb{E}\left[\hat{f}(x_0) - f(x_0)\right]^2$, $b(x_0) \stackrel{\text{def}}{=} \mathbb{E}\hat{f}(x_0) - f(x_0)$, $\sigma^2(x_0) = \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])^2]$. Then $MSE(x_0) = b^2(x_0) + \sigma^2(x_0)$ where all expectations are taken w.r.t a product distribution $\mathcal{D}^n$ defined on a sample of n points from the distribution $\mathcal{D}$.*

Lemma 1. *Consider the CAKE like density estimator as shown in equation (1) where α_{ij}'s are fixed positive real numbers such that $\sum_{j=1}^{m} \alpha_{ij} = 1 \; \forall i = 1, 2 \ldots n$. Denote by f_{max} the maximum value of the underlying density and $C_1 \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} k^2(\theta)\, \mathrm{d}\theta, C_2 \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} |\theta|^\beta k(\theta)\, \mathrm{d}\theta, C_3 \stackrel{\text{def}}{=} \frac{1}{n^2}(\frac{C_2 L}{l!})^2, C_4 \stackrel{\text{def}}{=} \frac{C_1 f_{max}}{n^2}$ and $|\cdot|$ is the standard Euclidean norm on $\mathbb{R}^d$. Under assumptions A1, A2 the estimator $\hat{f}$ has the following properties*

1) $\sigma^2(x_0) \leq \frac{C_1 f_{max}}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\alpha_{ij}^2}{h_j^d}$.
2) $|b(x_0)| \leq \frac{C_2 L}{n l!} \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} h_j^\beta$.
3) $MSE(x_0) \leq C_3 \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} h_j^\beta\right)^2 + C_4 \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\alpha_{ij}^2}{h_j^d} = \alpha^T M \alpha$, *where* $M \in R^{mn \times mn}, M \succ$

0 *is defined as*

$$M[ij,pl] = \begin{cases} C_3 h_j^{2\beta} + \frac{C_4}{h_j^d} & \textit{if } i = p \ \&j = l \\ C_3 h_j^{\beta} h_l^{\beta} & \textit{otherwise}. \end{cases} \tag{4}$$

Proof:

1) $\sigma^2(x_0) = \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^m \beta_{ij}\right]^2$, where $\beta_{ij} = \frac{\alpha_{ij}}{h_j^d} k\left(\frac{x-x_i}{h_j}\right) - \mathbb{E}\left[\frac{\alpha_{ij}}{h_j^d} k\left(\frac{x-x_i}{h_j}\right)\right]$. For given constants α_{ij} the r.v β_{ij} are independent with $\mathbb{E}[\beta_{ij}] = 0$. We have

$$\sigma^2(x_0) = \frac{1}{n^2}\mathbb{E}[\sum_{i=1}^n \sum_{j=1}^m \beta_{ij}]^2 \tag{5}$$

$$= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^m \mathbb{E}[\beta_{ij}^2] \text{ [Since } \beta_{ij} \text{ are 0 mean independent random variables]} \tag{6}$$

$$\leq \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^m \frac{\alpha_{ij}^2}{h_j^{2d}} \mathbb{E} k^2\left(\frac{x-x_i}{h_j}\right) \tag{7}$$

$$\leq \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^m \frac{\alpha_{ij}^2}{h_j^{d}} f_{\max} \int_{\mathbb{R}^d} k^2(\theta)\ \mathrm{d}\theta \tag{8}$$

$$= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^m \frac{\alpha_{ij}^2}{h_j^{d}} f_{\max} C_1. \tag{9}$$

2) To calculate $|b(x_0)|$ we first calculate the $\mathbb{E}\hat{f}(x_0)$. We have

$$\mathbb{E}\hat{f}(x_0) = \quad \tfrac{1}{n}\textstyle\sum_{i=1}^n \sum_{j=1}^m \mathbb{E}_{z\sim\mathcal{D}} \frac{\alpha_{ij}}{h_j^d} k\left(\frac{x_0-z}{h_j}\right) \tag{10}$$

$$= \quad \tfrac{1}{n}\textstyle\sum_{j=1}^m \frac{1}{h_j^d} \mathbb{E}_{z\sim\mathcal{D}} k\left(\frac{x_0-z}{h_j}\right) (\sum_{i=1}^n \alpha_{ij}) \tag{11}$$

$$= \quad \tfrac{1}{n}\textstyle\sum_{j=1}^m \int_{\mathbb{R}^d} k(\theta) f(x_0 + h_j\theta) s_j, \tag{12}$$

where $s_j \stackrel{\text{def}}{=} \sum_{i=1}^n \alpha_{ij}, \sum_{j=1}^m s_j = n, s_j \leq n\ \forall j = 1, 2 \ldots m$. Taylor expanding around x_0 and using the fact that all kernels are symmetric and of order $l = \lfloor \beta \rfloor$ we get

$$\left|E\hat{f}(x_0) - f(x_0)\right| = \quad \left| \tfrac{1}{n}\textstyle\sum_{j=1}^m \int_{\mathbb{R}^d} \frac{1}{l!} D^l f(x_0 + \tau h_j\theta) [\underbrace{h_j\theta, \ldots, h_j\theta}_{l \text{ times}}] s_j k(\theta)\ \mathrm{d}\theta \right| \tag{13}$$

$$\leq \quad \tfrac{1}{n}\textstyle\sum_{j=1}^m \int_{\mathbb{R}^d} \frac{L}{l!} h_j^{\beta} |\theta|^{\beta} s_j k(\theta)\ \mathrm{d}\theta \tag{14}$$

$$= \quad \tfrac{1}{n}\textstyle\sum_{j=1}^m s_j h_j^{\beta} \int_{\mathbb{R}^d} |\theta|^{\beta} k(\theta)\ \mathrm{d}\theta \tag{15}$$

$$= \quad \tfrac{C_2 L}{nl!}\textstyle\sum_{i=1}^n \sum_{j=1}^m \alpha_{ij} h_j^{\beta}. \tag{16}$$

where Equation (14) follows from Equation (13) by using the facts that the kernel $k(\cdot)$ is of order $l = \lfloor \beta \rfloor$ and $f \in \Sigma(\beta, L)$.

3) This follows trivially by the bias-variance decomposition and parts (1, 2).

□

This finishes the first part of our analysis. The second part of our analysis is to investigate this upper bound on the MSE and its optimal value. We do this by establishing the equivalence of the following four optimization problems for sufficiently large $n \geq n_0(f_{\max}, \beta, d, L)$.

$$P1: \min_{\alpha \in R^{nm \times 1}} \alpha^T M \alpha \qquad P2: \min_{\alpha \in R^m} \alpha^T P \alpha$$

$$\text{subject to:} \sum_{j=1}^{m} \alpha_{ij} = 1 \ \forall i = 1, \ldots, n \qquad \text{subject to:} \sum_{j=1}^{m} \alpha_j = 1$$

$$\alpha_{ij} \geq 0 \ \forall i = 1, \ldots, n, j = 1 \ldots, m \qquad \alpha_j \geq 0 \ \forall j = 1 \ldots m$$

$$P3: \min_{\alpha \in R^m} \alpha^T P \alpha \qquad P4: \min_{\alpha \in R^{nm \times 1}} \alpha^T M \alpha$$

$$\text{subject to:} \sum_{j=1}^{m} \alpha_j = 1 \qquad \text{subject to:} \sum_{j=1}^{m} \alpha_{ij} = 1 \ \forall i = 1, \ldots n.$$

$$v \stackrel{\text{def}}{=} \left[\sqrt{C_3} h_1^{\beta}, \ldots \sqrt{C_3} h_m^{\beta}\right]^T, D \stackrel{\text{def}}{=} \left[\frac{C_4}{h_1^d}, \ldots, \frac{C_4}{h_m^d}\right], P \stackrel{\text{def}}{=} vv^T + D. \tag{17}$$

Lemma 4. *The optimization problems P1 and P2 are equivalent to each other.*

Proof: The structure of the matrix M ensures that $\alpha_{i_1 j} = \alpha_{i_2 j} \ \forall i_1, i_2 = 1, 2 \ldots n$ and hence $\alpha^T M \alpha = n^2 \alpha^T P \alpha$. Hence optimization problems P1 and P2 are equivalent. □

The proof of next lemma follows trivially by using the Lagrangian of P3 and hence is omitted.

Lemma 5. *The solution to the optimization problem P3 is* $\alpha = \frac{P^{-1}\mathbf{1}_m}{\mathbf{1}_m P^{-1} \mathbf{1}_m}$.

Lemma 6. *Under assumption A1 and* $\forall n \geq n_0(f_{max}, \beta, d, L)$ *we have* $\alpha = \frac{P^{-1}\mathbf{1}_m}{\mathbf{1}_m^T P^{-1} \mathbf{1}_m} \geq 0$

Proof: Since $P = vv^T + D$ by Sherman-Morrison-Woodbury formula we will have

$$P^{-1} = D^{-1} - \frac{D^{-1} v v^T D^{-1}}{1 + v^T D^{-1} v}. \tag{18}$$

From Equation (17) we have $D^{-1} = \left[\frac{h_1^d}{C_4}, \ldots, \frac{h_m^d}{C_4}\right]$. Using Equation (17) we get

$$(vv^T)_{i,j} = C_3 h_i^\beta h_j^\beta \ \forall i,j = 1 \ldots m \tag{19}$$

$$(D^{-1}vv^TD^{-1})_{i,j} = \frac{C_3}{C_4^2} h_i^{\beta+d} h_j^{\beta+d} \ \forall i,j = 1, \ldots m \tag{20}$$

$$1 + v^T D^{-1} v = 1 + \frac{C_3}{C_4}\left(h_1^{2\beta+d} + h_2^{2\beta+d} + \ldots + h_m^{2\beta+d}\right) \tag{21}$$

Under assumption A1 and for large enough $n \geq n_0(f_{max}, \beta, d, L)$ such that $h_j \ll 1$ we have $1 + v^T D^{-1} v \approx 1$. Using Equations (18,19,20,21) we get

$$P_{i,j}^{-1} = \begin{cases} \frac{h_j^d}{C_4} - \frac{C_3}{C_4^2} h_j^{2\beta+2d} & \text{if } i = j \\ -\frac{C_3}{C_4^2} h_i^{\beta+d} h_j^{\beta+d} & \text{otherwise} \end{cases} \tag{22}$$

Using Equation (22) and the expression for α we get

$$\alpha_t = \frac{\frac{h_t^d}{C_4} - \frac{C_3}{C_4^2}\sum_{j=1}^m h_t^{\beta+d} h_j^{\beta+d}}{\sum_{i=1}^m \frac{h_j^d}{C_4} - \frac{C_3}{C_4^2}\sum_{i=1}^m \sum_{j=1}^m h_t^{\beta+d} h_j^{\beta+d}} \ \forall t = 1 \ldots m. \tag{23}$$

From Assumption A1 and for sufficiently large $n \geq n_0(f_{max}, \beta, d, L)$ we have $h_j^d >> h_j^{2\beta+2d}$. We get

$$\alpha_t \approx \frac{h_t^d}{\sum_{i=1}^m h_i^d} > 0 \ \forall t = 1, \ldots m.$$

□

Lemma 7. *Under assumption A1 and for large enough $n \geq n_0(f_{max}, \beta, d, L)$ the optimization problems P2 and P3 are equivalent.*

Proof: The difference between the optimization problems P2 and P3 is the absence of the positivity constraint on the alpha vector. However Lemma (6) guarantees that under the assumption A1, the resulting solution of optimization problem P3 is positive. Hence we can conclude that optimization problems P2 and P3 are equivalent. □

Lemma 8. *Problems P3 and P4 are equivalent and hence under assumption A1 and for large enough $n \geq n_0(f_{max}, \beta, d, L)$ P1 and P4 are equivalent.*

Proof: The proof for the first part of the Lemma is exactly the same as that for Lemma (4). The second part of the Lemma follows using lemmas (4), (7) and from the first part. □

The proof of next lemma is trivial and hence is omitted.

Lemma 9. *The solution of the optimization problem P4 is given by* $\alpha = M^{-1}A^T(AM^{-1}A^T)^{-1}\mathbf{1}_n$*, where* $A \in \mathbb{R}^{n\times nm}$ *and the* r^{th} *row of matrix* A *is given by the vector* $[\underbrace{\mathbf{0}_m,\ldots,\mathbf{0}_m}_{r-1\ times},\mathbf{1}_m,\underbrace{\mathbf{0}_m,\ldots\mathbf{0}_m}_{n-r\ times}]^T$.

Lemma 10. *Under A1, A2 and for large enough* $n \geq n_0(f_{max},\beta,d,L)$ *the optimal* $MSE(x_0) \leq \mathbf{1}_n^T B\mathbf{1}_n$ *where* $B = (AM^{-1}A^T)^{-1}$.

Proof: From Lemma (1) we know that $MSE(x_0) \leq \alpha^T M\alpha$. Now using Lemma (9) we get the required result. □

The final part of the proof is to analyze the magnitude of the optimal upper bound on $MSE(x_0)$. In order to be able to do this we need some simple lemmas.

Lemma 11. $\left|\mathbf{1}_n^T B^{-1}\mathbf{1}_n\right| \leq n\max\{|\lambda| : \lambda$ *is an eigen value of* $B\}$.

Proof: Matrix B is symmetric and hence is normal. Now for any vector x and normal matrix B we have $\left|\mathbf{1}_n^T B^{-1}\mathbf{1}_n\right| \leq n\max\{|\lambda| : \lambda$ is an eigen value of $B\}$ [2]. Simply Replace x with the vector $\mathbf{1}_n$ to get the desired result. □

Lemma 12. $\lambda_{min}(AM^{-1}A^T) \geq m\lambda_{min}(M^{-1})$.

Proof: Using the variational characterization of eigen values we have

$$\lambda_{\min}(AM^{-1}A^T) = \min_x \frac{(A^Tx)^TM^{-1}(A^Tx)}{x^Tx} \geq \frac{\lambda_{\min}(M^{-1})||A^Tx||_2^2}{x^Tx} = m\lambda_{\min}(M^{-1}) > 0$$

where the last inequality follows from the fact that $M \succ 0$. □

Lemma 13. $\lambda_{max}(M) \leq n\sum_{j=1}^m C_3h_j^{2\beta} + \frac{C_4}{h_j^d}$.

Proof: It is easy to see that $M = v_1v_1^T + D_1$ where $v_1 = [\underbrace{v, v\ldots v}_{ntimes}]$ and $D_1 = diag(\underbrace{D,\ldots D}_{ntimes})$ and where v, D are given by the Equation (17). Hence $\lambda_{max}(M) \leq \lambda_{max}(v_1v_1^T) + \lambda_{max}(D_1)$. Since $v_1v_1^T \succeq 0$, we get $\lambda_{max}(v_1v_1^T) \leq \sum_i \lambda_i(v_1v_1^T)$. Now $\sum_i \lambda_i(v_1v_1^T) = trace(v_1v_1^T) = n\sum_{i=1}^m C_3h_j^{2\beta}$. Since D is a diagonal matrix we get $\lambda_{max}(D) \leq \sum_{j=1}^m \frac{C_4}{h_j^d}$. □

Lemma 14. *If* $h_j = \Theta(n^{-\frac{1}{2\beta+d}})$ *then* $\forall n \geq n_0(f_{max}\beta,d,L)$ *under A1, A2 the optimal value of* $MSE(x_0)$ *is* $O(n^{-\frac{2\beta}{2\beta+d}})$.

Proof: We have the following chain of inequalities (where $B \stackrel{\text{def}}{=} (AM^{-1}A^T)^{-1}$).

$$\begin{aligned}
\mathbf{1}_n^T B \mathbf{1}_n \leq & \quad n \max\{|\lambda| : \lambda \text{ is an eigen value of B}\} && (24)\\
\leq & \quad n|||B|||_2 && (25)\\
= & \quad n \max\{\sqrt{\lambda} : \lambda \text{ is an eigen value of} B^2\} && (26)\\
= & \quad \tfrac{n}{\lambda_{min}(AM^{-1}A^T)} && (27)\\
\leq & \quad \tfrac{n}{m\lambda_{min}(M^{-1})} && (28)\\
= & \quad \tfrac{n\lambda_{max}(M)}{m} && (29)\\
\leq & \quad \tfrac{n^2}{m}\sum_{j=1}^{m} C_3 h_j^{2\beta} + \tfrac{n}{m}\sum_{j=1}^{m}\tfrac{C_4}{h_j^d} = O(n^{-\frac{2\beta}{2\beta+d}}) && (30)
\end{aligned}$$

Here Equation (25) follows from Equation (24) by the definition of spectral norm and the fact that spectral norm of a matrix is the smallest of all matrix norms. Equation (26) follows from Equation (25) by the definition of spectral radius, Equation (27) from Equation (26) from the definition of matrix B and the fact that matrix B is symmetric positive definite. Equation (28) follows from Equation (27) by lemma (12) and Equation (30) from Equation (29) using lemma (13). □

Lemma 2. *Consider the optimization problem*

$$\begin{aligned}
P1 : & \min_{\alpha \in R^{nm\times 1}} \alpha^T M \alpha \\
& \textit{subject to} : \sum_{j=1}^{m} \alpha_{ij} = 1 \ \forall i = 1, \ldots, n, \quad \alpha \geq 0.
\end{aligned}$$

Under assumptions A1, A2 and for $n \geq n_0(f_{max}, \beta, d, L)$ the optimal value of the objective is $\mathbf{1}_n^T(AM^{-1}A^T)^{-1}\mathbf{1}_n$, where $A \in R^{n\times nm}$ and the r^{th} row of the matrix A is given by $[\underbrace{\mathbf{0}_m, \ldots, \mathbf{0}_m}_{r-1 \ times}, \mathbf{1}_m, \underbrace{\mathbf{0}_m, \ldots \mathbf{0}_m}_{nm-r \ times}]^T$. Also the optimal value of $MSE(x_0) = O(n^{-\frac{2\beta}{2\beta+d}})$ is attained when $h_j = \Theta(n^{-\frac{1}{2\beta+d}})$.

Proof: The first part of the proof follows from Lemma (11). The second part of the proof follows from Lemma (15).

We are now ready to state and prove our main result.

Theorem 3. *Under assumptions A1, A2 and for* $h_j = \Theta(n^{-\frac{1}{2\beta+d}})\ \forall j = 1, \ldots, m$ *we have* $\forall n \geq n_1(C_2, L, l, \beta, d, L, K_{max})$ *the CAKE like estimator* $\hat{f}$ *given by Equation (1) satisfies*

$$\sup_{x_0 \in \mathbb{R}^d} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}_{\mathcal{D}^n} \left[(\hat{f}(x_0) - f(x_0))^2 \right] = O(n^{-\frac{2\beta}{2\beta+d}}). \tag{31}$$

Proof: It is enough to prove that $\forall f \in \Sigma(\beta, L), x_0 \in \mathbb{R}^d : f_{\text{max}} < C < \infty$, for some universal constant C. Then the result follows from Lemma (2). Choose a set of bounded smoothing kernels with $h_j = 1$. Now we have from the Lemma (1) that

$$f(x_0) \leq \frac{C_2 L}{l!} + \int K(x - z) f(z)\ \mathrm{d}z \leq \frac{C_2 L}{l!} + K_{\text{max}} < \infty. \tag{32}$$

Since the R.H.S. is independent of f, x_0, one can choose C to be the R.H.S. of the above equation. □

II. Risk of the CAKE estimator

In this section we shall prove an upper bound on the L_1 risk of the CAKE estimator in terms of its empirical risk via uniform stability arguments.

Theorem 4. *Let* $\hat{f}(x)$ *be the CAKE density estimator defined by the equation*

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\alpha_{ij}^*}{h_j^d} K\Big(\frac{x - x_i}{h_j}\Big) \tag{33}$$

where α^* *is the solution to the optimization problem*

$$T : \min_{\alpha}\ \alpha^T Z \alpha - 2\alpha^T v + \lambda ||\alpha||_2^2 \tag{34}$$

$$\textit{subject to: } \sum_{j=1}^{m} \alpha_{ij} = 1 \quad \forall i = 1, \ldots n, \alpha \geq 0 \tag{35}$$

where $Z \in \mathbb{R}^{nm \times nm}, \alpha, v \in \mathbb{R}^{nm \times 1}$ *are defined as*

$$Z[ij, pl] = \int \frac{1}{n^2 h_j^d h_l^d} k\left(\frac{x - x_i}{h_j}\right) k\left(\frac{x - x_p}{h_l}\right)$$

$$v[ij] = \frac{1}{n^2} \sum_{\substack{p=1 \\ p \neq i}}^{n} \frac{1}{h_j^d} k\left(\frac{x_i - x_p}{h_j}\right). \tag{36}$$

Suppose we are provided fixed bandwidths $h_1, \ldots, h_m$ *and the true underlying density function* f *is bounded by a constant* B*. Let* $c_d = (\sqrt{2\pi})^d$*. Then with probability atleast* $1 - \delta$ *over the input*

training samples, the CAKE estimator $\hat{f}$ with Gaussian base kernels satisfies the risk bound

$$\mathbb{E}_{x\sim\mathcal{D}}|\hat{f}(x)-f(x)| \leq \frac{1}{n}\sum_{i=1}^{n}|\hat{f}(x_i)-f(x_i)|+$$
$$\left[\frac{4}{c_d}\Big[\sum_{j=1}^{m}\frac{1}{h_j^d}+\frac{2\sqrt{2}}{\sqrt{c_d}\sqrt{\lambda}}\sqrt{\sum_{j=1}^{m}\frac{1}{h_j^{2d}}}\sqrt{\sum_{j=1}^{m}\sum_{l=1}^{m}\frac{1}{(\sqrt{h_j^2+h_l^2})^d}}\Big]+B+\sum_{j=1}^{m}\frac{1}{c_d h_j^d}\right]\sqrt{\frac{ln(\frac{1}{\delta})}{2n}}. \quad (37)$$

Before we prove the theorem we need the definition of uniform stability of a learning algorithm. In the definition to follow we denote by A any learning algorithm. The learning algorithm A learns on S and outputs a function (model) A_S. Also let $A_{S_{-i}}$ denote the function outputted by learning algorithm A when trained on the dataset S_{-i}. Uniform stability quantifies the algorithms behaviour when any arbitrary point from a set S is not used for training.

Definition 3. *An algorithm A has uniform stability β w.r.t the loss function l if the following holds true*

$$\forall S=\{x_1,\ldots,x_n\},\forall i\in\{1,\ldots,n\}: ||l(A_S,\cdot)-l(A_{S_{-i}},\cdot)||_\infty \leq \beta \quad (38)$$

The proof of the above theorem proceeds by bounding the uniform stability of the CAKE algorithm. The following is an important result concerning the expected risk of a learning algorithm in terms of its empirical risk.

Theorem 5. *Let A be an algorithm with uniform stability β w.r.t a loss function l such that the function A_S learnt by the algorithm A when trained on dataset S satisfies $\forall x, S: 0\leq l(A_S,x)\leq M$. Then for any $n\geq 1$, and any $\delta\in(0,1)$ with probability atleast $1-\delta$ over the random draw of the sample S we have*

$$\mathbb{E}_{x\sim\mathcal{D}}\ l(A_s,x) \leq \frac{1}{n}\sum_{i=1}^{n}l(A_s,x_i)+2\beta+(4n\beta+M)\sqrt{\frac{log(1/\delta)}{2n}}. \quad (39)$$

Theorem (5) and definition (3) were first provided by Bousquet and Elisseeff ([3]) in the context of supervised learning problems. However it is straightforward to see that their result holds true for unsupervised learning problems also.

Lemma 6. *Given m fixed bandwidths $h_1,\ldots,h_m$, the uniform stability β of the CAKE estimator obtained by solving the optimization problem T (equations (34-35)) w.r.t the loss function*

$l(A_S, x) = |\hat{f}(x) - f(x)|$ *with Gaussian base kernels is upper bounded by*

$$\beta \leq \frac{2}{n\sqrt{\lambda}(\sqrt{2\pi})^{\frac{3d}{4}}} \sqrt{\sum_{j=1}^{m} \frac{1}{h_j^{2d}}} \sqrt{2\sum_{l=1}^{m}\sum_{j=1}^{m} \frac{1}{\left(\sqrt{h_j^2 + h_l^2}\right)^d}} + \frac{1}{n(\sqrt{2\pi})^d} \sum_{j=1}^{m} \frac{1}{h_j^d}. \tag{40}$$

Proof: Let $l(A_S, x) \stackrel{\text{def}}{=} |\hat{f}(x) - f(x)|$. Let $S_{-n} = \{x_1, \ldots, x_n\} - \{x_n\}$ represent the dataset which doesn't have the training point x_n. The deleted CAKE density estimate learnt using dataset S_{-n} is

$$\hat{f}_{-n}(x) \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_{ij}^*}{h_j^d} k\left(\frac{x - x_i}{h_j}\right) \tag{41}$$

where β^* is the solution to the optimization problem

$$V : \min_{\beta \in \mathbb{R}^{nm}} \beta^T M \beta - 2\beta^T v \tag{42}$$

$$\text{subject to: } \sum_{j=1}^{m} \beta_{ij} = 1 \qquad \forall i = 1, \ldots n, \beta \geq 0 \tag{43}$$

$$\beta_{nj} = 0 \ \forall j = 1, \ldots m \tag{44}$$

where $M = Z + \lambda I$ and Z is defined in equation (36). Note that the optimization problem T given by equations (34-35) and the problem V differ by an additional constraint in V given by the constraint (44). By definition bounding stability is equivalent to bounding $\forall i : |l(A_S, x) - l(A_{S_{-i}}, x)|$. We shall bound $|l(A_S, x) - l(A_{S_{-n}}, x)|$ and show that the bound doesn't depend on x_n, which allows us to get a bound on β. Let α^* denote the optimal solution of problem T and β^* the optimal solution of problem V.

$$\left||\hat{f}(x) - f(x)| - |\hat{f}_{-n}(x) - f(x)|\right| \leq |\hat{f}(x) - \hat{f}_{-n}(x)| \leq \underbrace{\frac{1}{n}\left|\sum_{i=1}^{n-1}\sum_{j=1}^{m} \frac{1}{h_j^d}(\alpha_{ij}^* - \beta_{ij}^*) k\left(\frac{x - x_i}{h_j}\right)\right|}_{T_1} + \underbrace{\frac{1}{n}\left|\sum_{j=1}^{m} \alpha_{nj}^* \frac{1}{h_j^d} k\left(\frac{x - x_n}{h_j}\right)\right|}_{T_2} \tag{45}$$

It is enough to upper bound T_1, T_2. Let us denote by $\Pi(\alpha^*)$ the vector obtained by setting the last m components of α^* to 0. Then by Cauchy-Scwartz inequality we have

$$T_1 \leq \frac{1}{n} ||\Pi(\alpha^*) - \beta^*|| \ ||K_v|| \tag{46}$$

where $K_v \in \mathbb{R}^{(n-1)m}$ and $K_v[ij] = \frac{1}{h_j^d} k\left(\frac{x-x_i}{h_j}\right)\ \forall i = 1, \ldots n-1, j = 1 \ldots m$. Since all our base kernels are Gaussian one can trivially upper bound

$$||K_v||_2 \leq \sqrt{\frac{n}{(\sqrt{2\pi})^{2d}} \sum_{j=1}^{m} \frac{1}{h_j^{2d}}}. \tag{47}$$

Hence now it is enough to upper bound the quantity $||\Pi(\alpha^*) - \beta^*||$. Let us denote the objective of optimization problem T by $g_T(\alpha)$ and that of problem V by $g_V(\beta)$. Note that $g_T(\alpha), g_V(\beta) = \infty$ if α, β do not belong to the feasible set of the optimization problems T and V respectively. For any vector $\theta \in \mathbb{R}^{nm}$ which is a feasible solution of the optimization problem T, define

$$\Pi(\theta) \stackrel{\text{def}}{=} [\theta_{11}, \ldots, \theta_{(n-1)m}, \underbrace{0, \ldots, 0}_{m}]. \tag{48}$$

i.e $\Pi(\theta)$ is the vector obtained by projecting θ onto the feasible set of optimization problem V. Observe that $g_T(\alpha), g_V(\beta)$ are both λ strongly convex in α, β respectively, and hence due to the strong convexity property and the fact that β^* is an optimal solution of optimization problem Q we get

$$g_V(\Pi(\alpha^*)) - g_V(\beta^*) \geq \frac{\lambda}{2} ||\Pi(\alpha^*) - \beta^*||^2. \tag{49}$$

By definition $g_T(\theta)$ is a quadratic function involving terms not having θ_{nj}'s and terms involving $\theta_{nj}, j = 1, \ldots, m$. Since the last m terms in $\Pi(\theta)$ are 0 by definition hence $g_V(\Pi(\theta))$ has all the terms as in $g_T(\theta)$ except for the terms containing θ_{nj}. Hence we have for any θ satisfying the constraints of P and $\Pi(\theta)$ satisfying the constraints of problem V

$$g_V(\Pi(\theta)) = g_T(\theta) - \sum_{j=1}^{m} T_{\theta_{nj}}. \tag{50}$$

Here $T_{\theta_{nj}}$ are terms in $g_T(\theta)$ involving θ_{nj}. Now for the vector β^* let $\bar{\beta}^*$ be defined as

$$\bar{\beta^*}_{ij} \stackrel{\text{def}}{=} \beta^*_{ij} \qquad \forall i = 1, \ldots, n-1, j = 1 \ldots, m \tag{51}$$

$$\bar{\beta^*}_{nj} \stackrel{\text{def}}{=} \alpha^*_{nj} \qquad \forall j = 1, \ldots m \tag{52}$$

Hence β^* belongs to the feasible set of problem V and $\bar{\beta}^*$ belongs to the feasible set of problem T. Using equation (53) we get

$$g_V(\Pi(\alpha^*)) - g_V(\beta^*) = \underbrace{g_T(\alpha^*) - g_T(\bar{\beta}^*)}_{\leq 0} + \sum_{j=1}^{m} T_{\bar{\beta^*}_{nj}} - \sum_{j=1}^{m} T_{\alpha^*_{nj}} \quad \leq \sum_{j=1}^{m} T_{\bar{\beta^*}_{nj}} - \sum_{j=1}^{m} T_{\alpha^*_{nj}} \tag{53}$$

where the last inequality follows from the fact that α^* is an optimal solution of the optimization problem T. From the definitions of the function $g_T(\cdot)$ and M and using equation (51) we get

$$\sum_{j=1}^{m} T_{\bar{\beta^*}_{nj}} - \sum_{j=1}^{m} T_{\alpha^*_{nj}} = 2\sum_{j=1}^{m}\alpha^*_{nj}\sum_{p=1}^{n-1}\sum_{l=1}^{m}(\beta^*_{pl}-\alpha^*_{pl})M[nj,pl] + \underbrace{\sum_{j=1}^{m}(\bar{\beta^*_{nj}}^2 - (\alpha^*_{nj})^2)M[nj,nj]}_{0}$$

$$+ \underbrace{2\sum_{j=1}^{m}\sum_{\substack{l=1\\ l\neq j}}^{m}(\bar{\beta^*}_{nj}\bar{\beta^*}_{nl} - \alpha_{nj}\alpha_{nl})M[nj,nl]}_{0} + \underbrace{2\sum_{j=1}^{m}(\bar{\beta^*}_{nj}-\alpha^*_{nj})v_{nj}}_{0}$$

$$\leq 2\max_{j=1\ldots m}\sum_{p=1}^{n-1}\sum_{l=1}^{m}(\beta^*_{pl}-\alpha^*_{pl})M[nj,pl] \leq n\sum_{j=1}^{m}\sum_{l=1}^{m}\frac{2}{n^2(\sqrt{2\pi})^d}\frac{1}{\left(\sqrt{h_j^2+h_l^2}\right)^d}. \quad (54)$$

where the penultimate inequality follows from the fact that α^*_{nj}'s form a convex combination and the last inequality from the definition of the matrix M. Now using equations (46,47,49,53,54) we get

$$T_1 \leq \frac{1}{n}\sqrt{\frac{1}{(\sqrt{2\pi})^{2d}}\sum_{j=1}^{m}\frac{1}{h_j^{2d}}}\sqrt{\frac{2}{\lambda}\sum_{j=1}^{m}\sum_{l=1}^{m}\frac{2}{(\sqrt{2\pi})^d}\frac{1}{\left(\sqrt{h_j^2+h_l^2}\right)^d}}. \quad (55)$$

Now all we need to do is upper bound T_2. We have

$$T_2 = \frac{1}{n}\left|\sum_{j=1}^{m}\frac{\alpha^*_{nj}}{h_j^d}k\Big(\frac{x-x_i}{h_j}\Big)\right| \leq \frac{1}{n(\sqrt{2\pi})^d}\sum_{j=1}^{m}\frac{1}{h_j^d}. \quad (56)$$

Putting together the bounds for T_1, T_2 we get

$$\beta \leq \frac{1}{n}\sqrt{\frac{1}{(\sqrt{2\pi})^{2d}}\sum_{j=1}^{m}\frac{1}{h_j^{2d}}}\sqrt{\frac{2}{\lambda}\sum_{j=1}^{m}\sum_{l=1}^{m}\frac{2}{(\sqrt{2\pi})^d}\frac{1}{\left(\sqrt{h_j^2+h_l^2}\right)^d}} + \frac{1}{n(\sqrt{2\pi})^d}\sum_{j=1}^{m}\frac{1}{h_j^d}. \quad (57)$$

□

The next lemma bounds the loss $l(A_S, x) = |\hat{f}(x) - f(x)|$.

Lemma 7. *Suppose the underlying density function is bounded by a universal constant B. Then the CAKE estimator with Gaussian kernels as base kernels and with fixed bandwidths $h_1, \ldots, h_m$ satisfies $|\hat{f}(x) - f(x)| \leq B + \frac{1}{(\sqrt{2\pi})^d}\sum_{j=1}^{m}\frac{1}{h_j^d}$.*

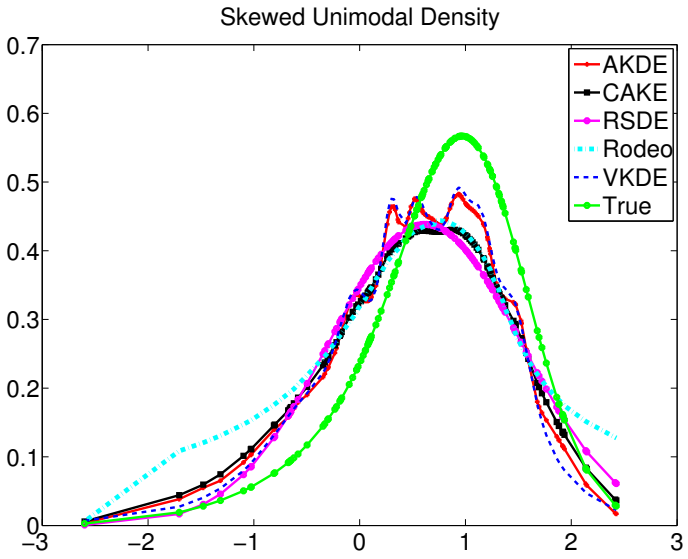
Proof: We have

$$|\hat{f}(x) - f(x)| \leq |\hat{f}(x)| + f(x) \leq B + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\alpha_{ij}}{h_j^d}k\Big(\frac{x-x_i}{h_j}\Big) \leq B + \frac{1}{(\sqrt{2\pi})^d}\sum_{j=1}^{m}\frac{1}{h_j^d}\square \quad (58)$$

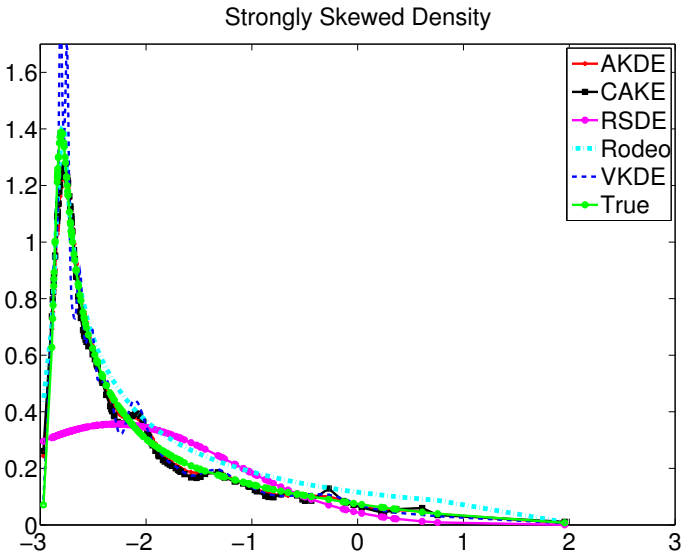
Proof of Theorem 4. Applying theorem (5) to the CAKE algorithm with the loss function $l(A_s, x) \stackrel{\text{def}}{=} |\hat{f}(x) - f(x)|$ and lemmas (6,7) we get the necessary result. □

III. Additional Experimental Results

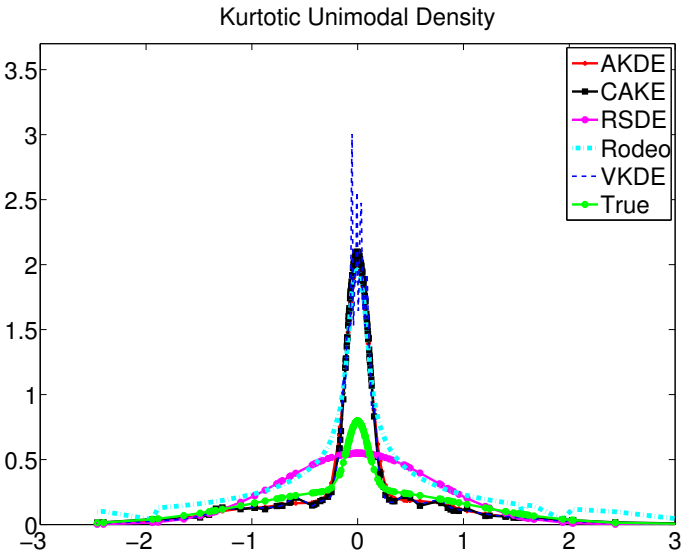
In the main paper we showed a couple of results comparing CAKE to other density estimators on some benchmark 1-d synthetic datasets. In this supplement we show the results on all the benchmark 1-d synthetic datasets. All the synthetic distributions are from Marron and Wand ([4]).



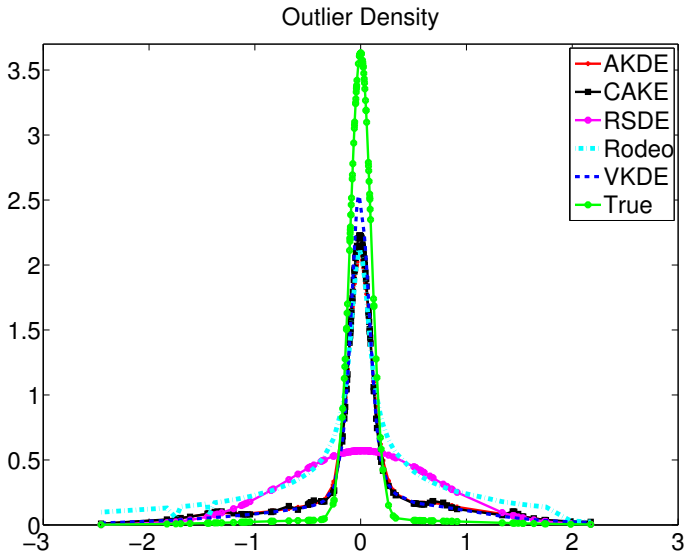
(a) Skewed Unimodal Density



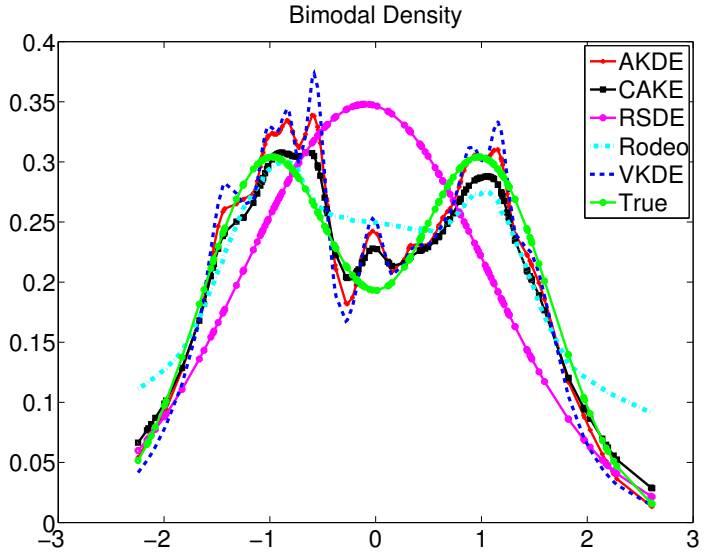
(b) Strongly Skewed Density



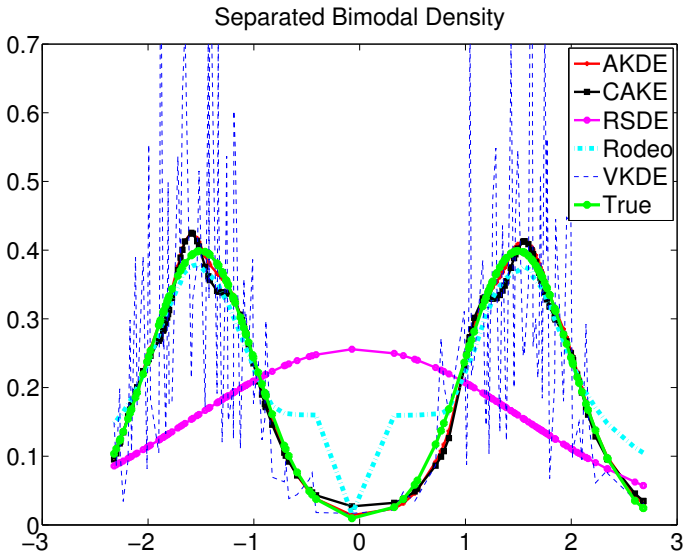
(c) Kurtotic Unimodal Density



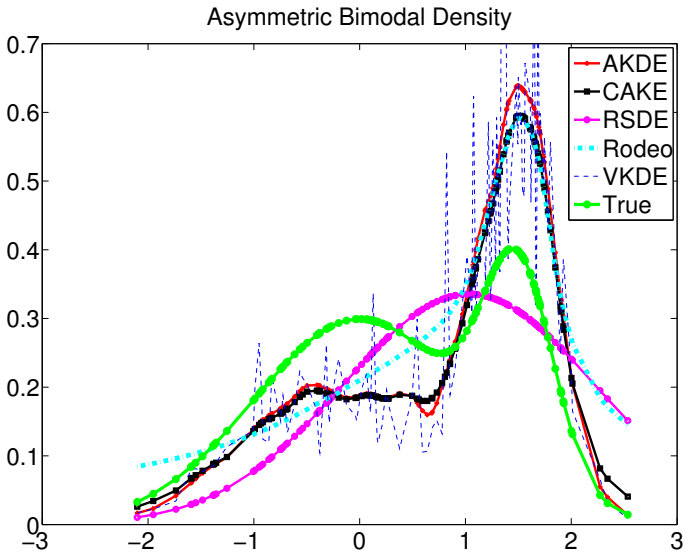
(d) Outlier Density



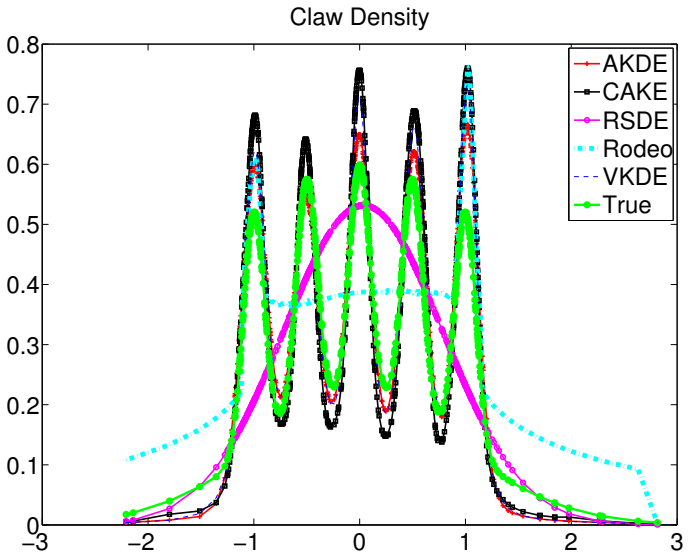
(e) Bimodal Density



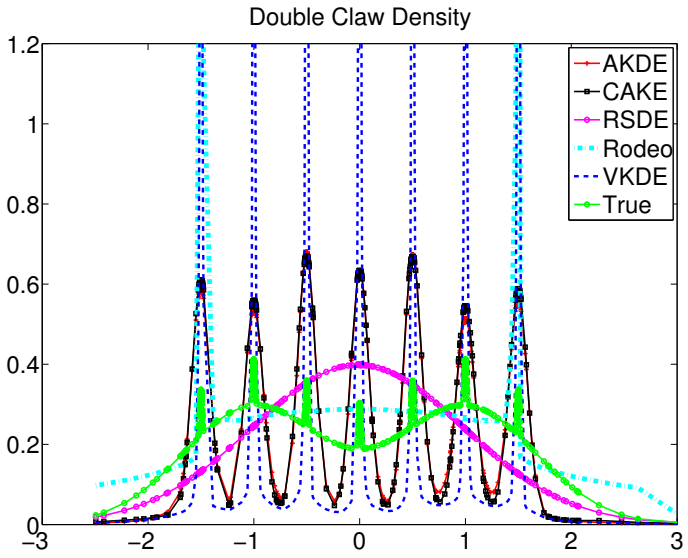
(f) Separated Bimodal Density



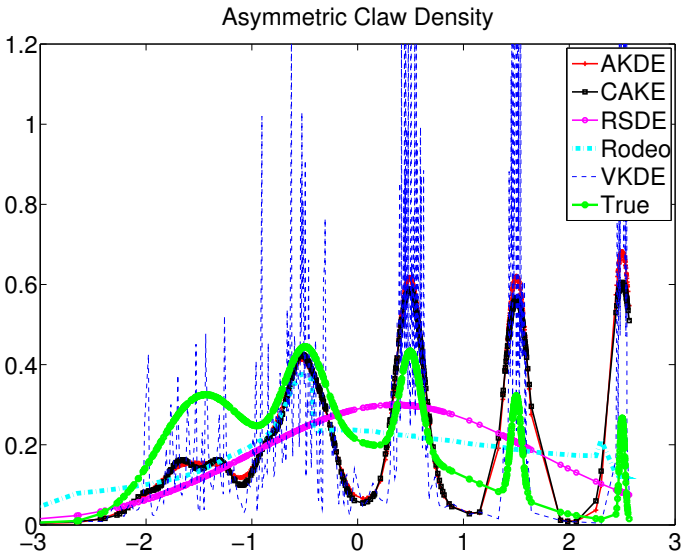
(g) Asymmetric Bimodal Density



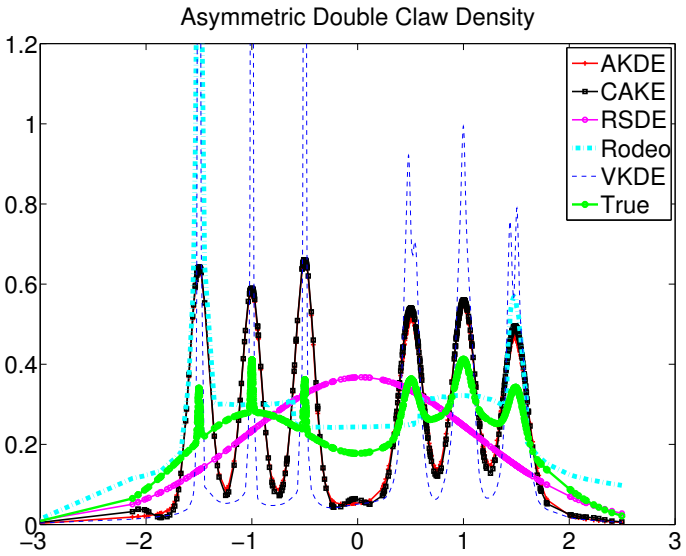
(h) Claw Density



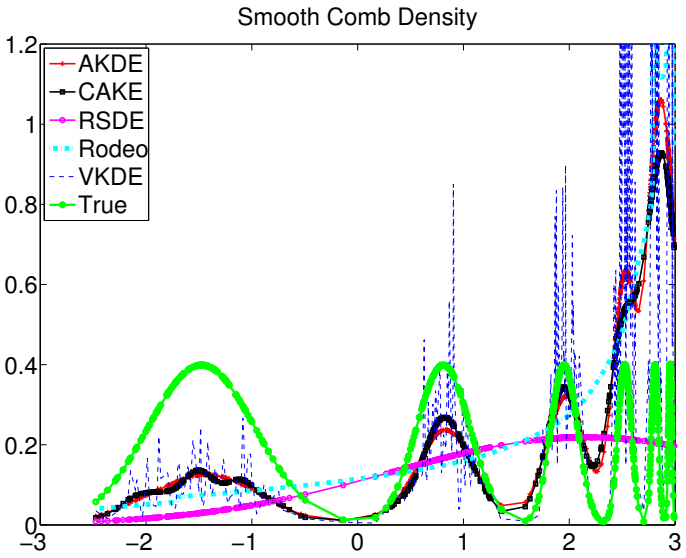
(i) Double Claw Density



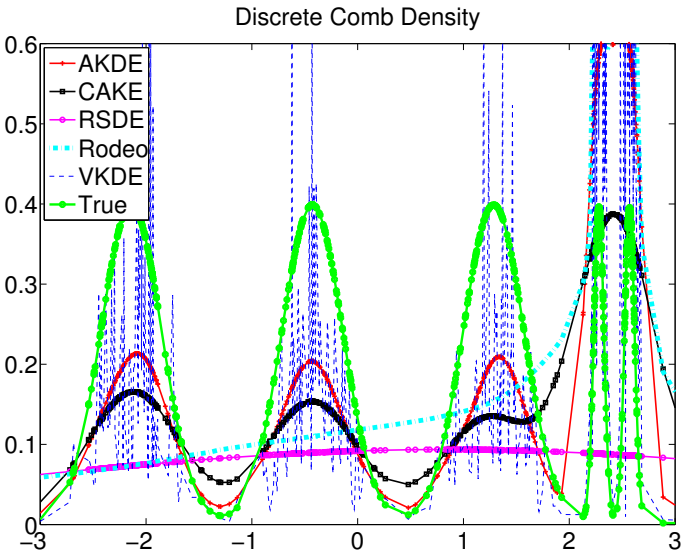
(j) Assymetric Claw Density



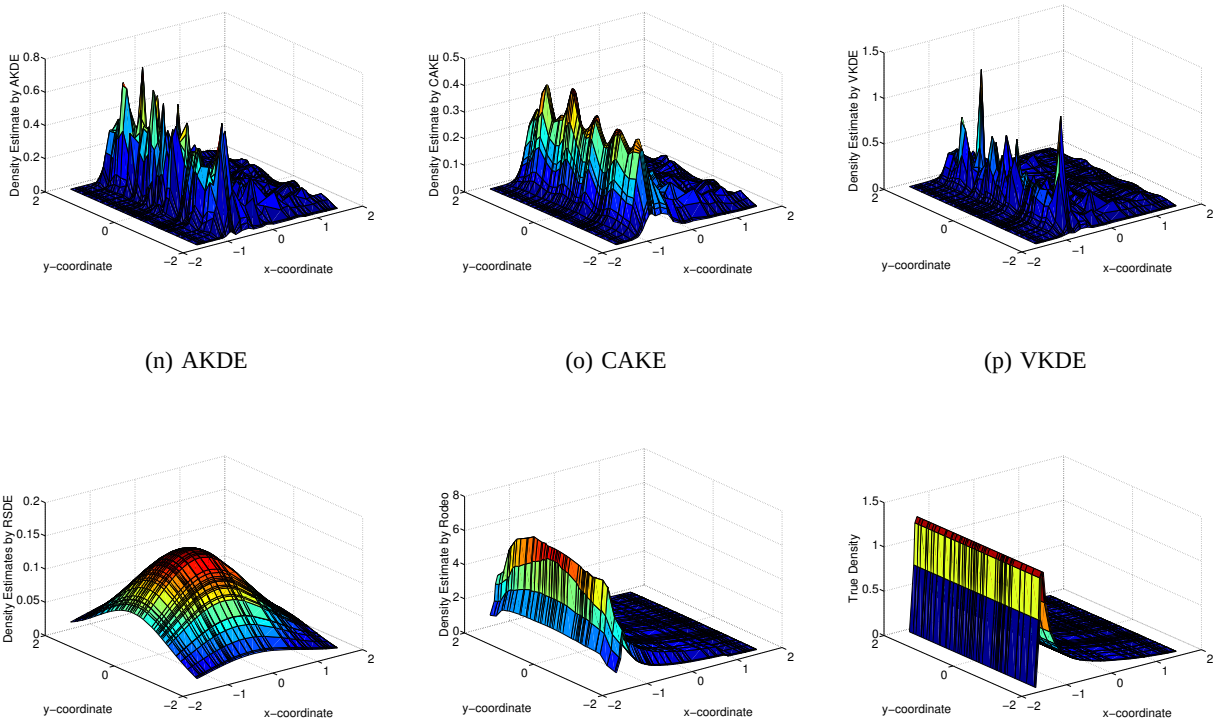
(k) Assymetric Double Claw Density



(l) Smooth Comb Density



(m) Discrete Comb Density

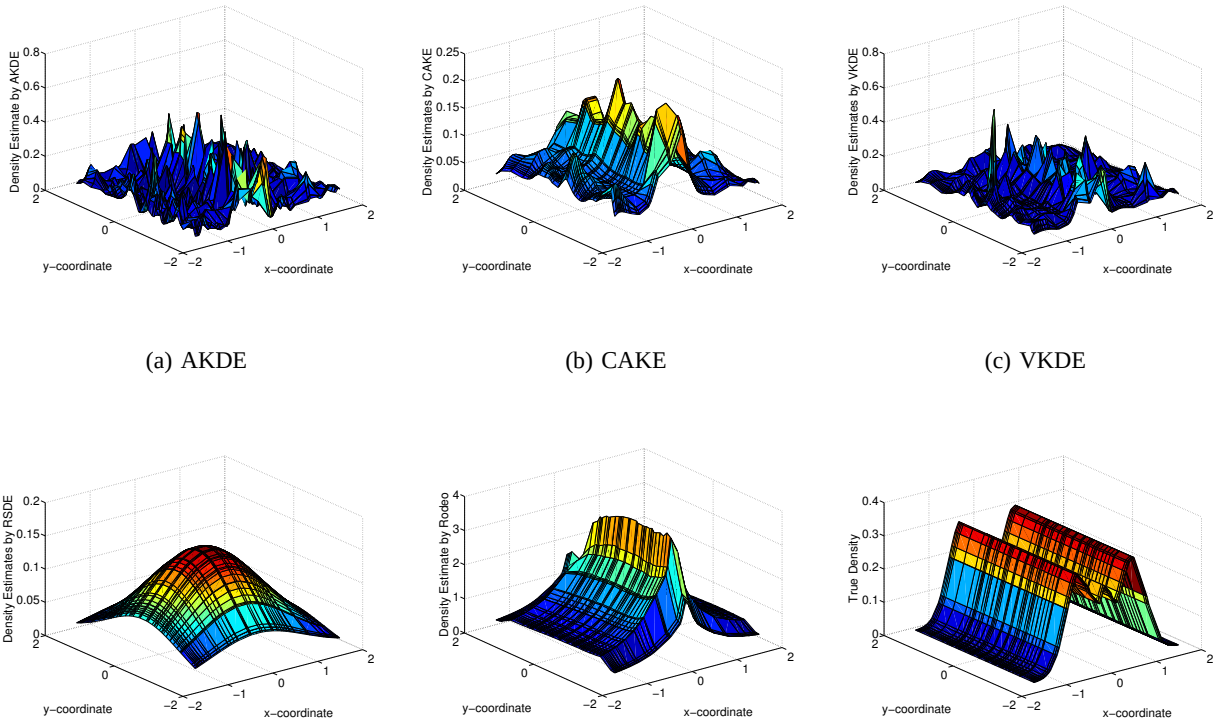


(n) AKDE (o) CAKE (p) VKDE

(q) RSDE (r) RODEO (s) True

Fig. 1: Performance on a product distribution with strongly skewed density along x-axis and unifrom density along y-axis.

figure

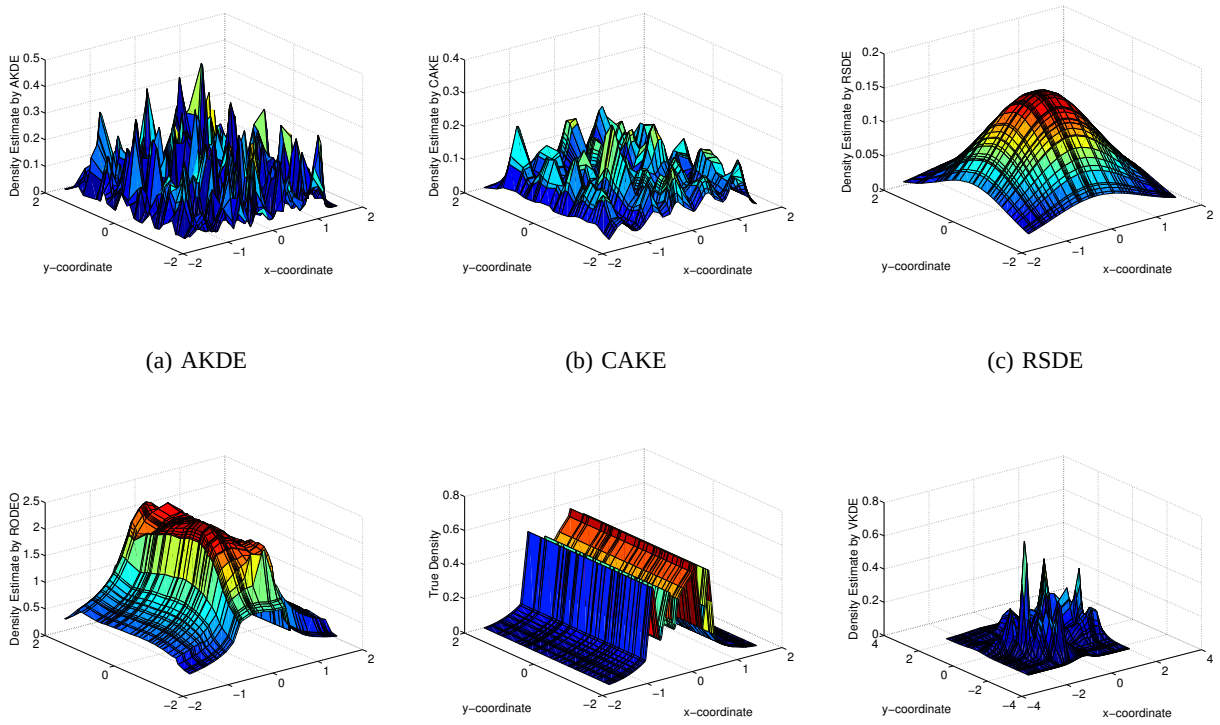


(a) AKDE (b) CAKE (c) VKDE

(d) RSDE (e) RODEO (f) True

Fig. 2: Performance of various density estimators on a product distribution with Trimodal density along x-axis and unifrom density along y-axis.

figure

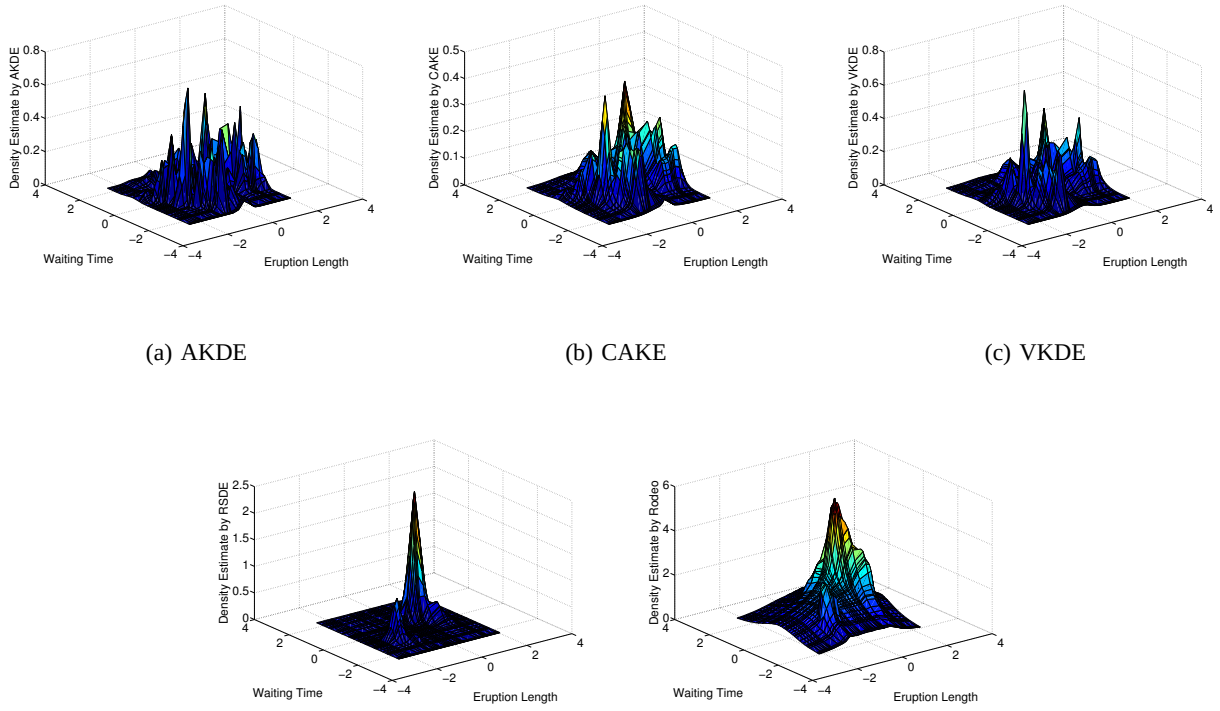


(a) AKDE (b) CAKE (c) RSDE

(d) RODEO (e) True (f) VKDE

Fig. 3: Performance of various density estimators on a product distribution with Claw density along x-axis and unifrom density along y-axis.

figure



(a) AKDE (b) CAKE (c) VKDE

(d) RSDE (e) RODEO

Fig. 4: Performance of various density estimators on the multi-dimensional version of old faithful dataset.

figure

Density function	**CAKE**	**Adaptive**	**Variable**	**RODEO**	**RSDE**
Skewed Unimodal	0.0960	0.0829	0.0827	0.0982	0.1061
Strongly Skewed	0.05907	0.06419	0.1305	0.0584	0.468
Kurtotic Unimodal	0.7688	0.721564	0.8326	0.7144	0.1604
Outlier Density	0.80	0.851	0.6907	0.8856	1.70
Bimodal Density	0.02073	0.02337	0.1979	0.0313	0.079
Separated Bimodal Density	0.02963	0.01272	0.1723	0.0377	0.168
Asymmetric Bimodal Density	0.1242	0.1467	0.1928	0.1258	0.0747
Claw	0.1030	0.0639	0.0781	0.1272	0.1597
Double claw	0.2683	0.2590	2.755	1.3159	0.1288
Asymmetric claw	0.2378	0.2449	0.9883	0.1194	0.1248
Asymmetric double claw	0.2097	0.2036	2.1711	1.0060	0.1199
Smooth Comb	0.3543	0.4039	1.5482	1.1866	0.175
Discrete comb	0.1489	0.1980	0.6887	0.4419	0.2092
Strongly skewed+Uniform	0.0939	0.38	0.3816	1.4232	0.3654
Claw+ Uniform	0.1797	0.1873	0.181	1.0837	0.1708
Trimodal+Uniform	0.1239	0.1390	0.1300	1.10067	0.1151

TABLE I: RMSE values of different density estimators on various synthetic 1-d and 2-d distribution.

table

REFERENCES

[1] A.B. Tsybakov. *Introduction to nonparametric estimation*. Springer Verlag, 2009.

[2] R.A. Horn and C.R. Johnson. *Matrix analysis*. Cambridge Univ Press, 1990.

[3] O. Bousquet and A. Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

[4] J.S. Marron and M.P. Wand. Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736, 1992.