RATES OF CONVERGENCE FOR THE GAUSSIAN MIXTURE SIEVE

Chris Genovese and Larry Wasserman[1]
Carnegie Mellon University
November 11, 1998

Gaussian mixtures provide a convenient method of density estimation that lies somewhere between parametric models and kernel density estimators. When the number of components of the mixture is allowed to increase as sample size increases, the model is called a mixture sieve. We establish a bound on the rate of convergence in Hellinger distance for density estimation using the Gaussian mixture sieve assuming that the true density is itself a mixture of Gaussians; the underlying mixing measure of the true density is not assumed to necessarily have finite support. Computing the rate involves some delicate calculations since the size of the sieve — as measured by bracketing entropy — and the saturation rate, cannot be found using standard methods. When the mixing measure has compact support, using $k_n \sim n^{2/3}/(\log n)^{1/3}$ components in the mixture yields a rate of order $(\log n)^{(1+\eta)/6}/n^{1/6}$ for every $\eta > 0$. When the mixing measure is not compactly supported, we find that the rates depend heavily on the tail behavior of the true density. The sensitivity to the tail behavior is diminished by using a robust sieve which includes a long-tailed component in the mixture. Then, a spectrum of interesting rates arise depending on the thickness of the tails of the mixing measure.

KEY WORDS: Density Estimation; Mixtures, Rates of Convergence, Sieves.

1. INTRODUCTION

Statistical inference using mixtures of Gaussians is used for many purposes including density estimation, clustering and robust estimation; see, for example, Lindsay (1995), McLachlan and Basford (1988), Banfield and Raftery (1993) and Robert (1996). When the number of components of the mixture is allowed to increase with sample size, the model is called a Gaussian mixture sieve (Grenander, 1982, Wong and Shen 1995). These sieves have been studied by several authors including Geman and Hwang (1982), Roeder (1992), Priebe (1994) and Roeder and Wasserman (1997). Related work from a Bayesian point of view is discussed in Escobar and West (1995). Priebe argues that in many cases, a mixture sieve has many advantages as a density estimator over kernel density estimates. For example, Priebe showed that with $n = 10,000$ observations, a log-normal density can be well approximated by a mixture of about 30 normals. In contrast, a kernel density estimator uses a mixture of

[1]Genovese's research supported by NSF grant DMS 9705034. Wasserman's research supported by N.I.H. grant R01-CA54852-07 and NSF grant DMS-9803433. The authors thank Kert Viele for helpful discussions.

10,000 normals. Despite the ubiquity of mixture sieve models, little is known about their asymptotic properties. In particular, the rate of convergence of the density estimator of this sieve has not been established. In this paper, we bound the rate of convergence. (Rates of convergence for the mixing distribution function have been studied in Chen (1995).)

Let $\phi(x;\mu,\sigma)$ denote a Gaussian density with mean μ and variance σ^2. A finite Gaussian mixture is a density of the form

$$f_\theta(x) = \sum_{j=1}^{k} p_j \phi(x;\mu_j,\sigma_j) \tag{1}$$

where $\theta = (\mu,\sigma,p)$, $\mu = (\mu_1,\ldots,\mu_k)$, $\sigma = (\sigma_1,\ldots,\sigma_k)$, $p = (p_1,\ldots,p_k)$. Here, the μ_j's are real, the σ_j's are positive reals, $p_j \geq 0$ and $\sum_j p_j = 1$.

Let m_k, s_k and S be positive constants such that $m_k \to \infty$ and $s_k \downarrow 0$ as $k \to \infty$ and let

$$\mathcal{M}_k = \left\{ f(\cdot) = \sum_{j=1}^{k} p_j \phi(\cdot\,;\mu_j,\sigma_j);\ |\mu_j| \leq m_k,\ \text{and } s_k \leq \sigma_j \leq S,\ j=1,\ldots,k \right\}. \tag{2}$$

Let k_n be a sequence of integers such that $k_n \to \infty$ as $n \to \infty$. The sieve we are interested in is $\mathcal{M}_{k_n}$. Our estimate of the true density is $\hat{f}_n(\cdot) = f_{\hat{\theta}}(\cdot)$ where $\hat{\theta}$ is the maximum likelihood estimate of θ in the model (2). We have chosen to fix S mainly for convenience. This parameter can also be allowed to increase with k but the results do not change materially.

We will assume that the true density is a "general Gaussian mixture" of the form

$$f_0(x) = \int_0^\infty \int_{-\infty}^{\infty} \phi(x;\mu,\sigma)\, dP(\mu,\sigma) \tag{3}$$

for some probability measure P on the Borel σ-algebra over $\mathcal{R} \times \mathcal{R}^+$. Let $\mathcal{F}$ denote all such densities. Of course, $\mathcal{F}$ contains all finite and countable mixtures as a special case. It is worth noting that the Dirichlet process mixture prior used in nonparametric Bayesian inference (Escobar and West 1995) uses a prior with support in $\mathcal{F}$.

We measure the error in the estimate by Hellinger distance $d_H(f_0,\hat{f}_n)$ where $d_H(f,g)^2 = \int(\sqrt{f}-\sqrt{g})^2$. We bound the rate at which $d_H(f_0,\hat{f}_n)$ goes to 0, in two steps: (i) we bound the likelihood ratio outside Hellinger neighborhoods of the true density and (ii) we compute the rate at which finite mixtures saturate the set $\mathcal{F}$. The first task is addressed in Section 2 by computing the "size" of $\mathcal{M}_{k_n}$ in terms of its Hellinger bracketing entropy, and then appealing to recent results of Wong and Shen (1995). The second task is addressed in Section 3. Calculating the bracketing entropy and the saturation rate is usually straightforward for finite dimensional models. However, mixture models do not behave as nicely as most finite dimensional parametric families so these calculations require special attention. In particular, the square root of the density of a mixture model is not differentiable everywhere so that standard methods for computing entropy are not available. Hence, we believe that the calculations in Sections 2 and 3 might be useful for other non-regular families as well. We put these pieces together and compute the rate in Section 4. Specifically we find t_n such that $P^*(d_H(f_0, \hat{f}_n) > t_n) = o(1)$. In Section 5 we discuss an improvement on the sieve. Section 6, contains closing remarks and unsolved problems.

The main conclusion of this paper is as follows. If the mixing measure P is compactly supported, then taking $k_n \sim n^{2/3}(\log n)^{2/3}$ yields the rate $(\log n)^{(1+\eta)/6}/n^{1/6}$ for every $\eta > 0$. If the mixing measure is not compactly supported then we cannot compute the rate without the adjusted sieve in Section 5. In this case we get a spectrum of rates depending on the tail behavior of P.

2. BOUNDING THE LIKELIHOOD RATIO

In this section we bound the supremum of the likelihood outside a Hellinger neighborhood of the true density. Throughout, we consider densities on the real line with respect to Lebesgue measure. Let $d_H(f,g)$ be Hellinger distance, $d_{TV}(f,g)$ total variation distance and $D(f,g)$ be Kullback-Leibler divergence, i.e.

$$\begin{aligned} d_H^2(f,g) &= \int (\sqrt{f}(x) - \sqrt{g}(x))^2 dx = \int f(x)dx + \int g(x)dx - 2\int \sqrt{f(x)g(x)}dx \\ d_{TV}(f,g) &= \frac{1}{2}\int |f(x) - g(x)|dx \\ D(f,g) &= \int f(x)\log f(x)/g(x)dx. \end{aligned}$$

The following inequalities are well known and will be used in what follows.

PROPOSITION. *Consider integrable functions f and g, not necessarily probability density functions and suppose that $g(x) \le f(x)$ for almost all x with respect to Lebesgue measure. Then, $d_H(f,g) \le \sqrt{2d_{TV}(f,g)}$ and $d_{TV}(f,g) \le d_H(f,g)\sqrt{\int f(x)dx}$.*

2.1. The Bracketing Entropy of $\mathcal{M}_k$. In this subsection we measure the size of $\mathcal{M}_k$ using bracketing entropy (van der Vaart and Wellner 1997, section 2.7). If $\mathcal{A}$ is a set non-negative, integrable functions and d is a metric on this set, then an ϵ-bracketing (with respect to d) is a set of pairs of integrable functions $(\ell_1, u_1), \ldots, (\ell_m, u_m)$ such that

(i) for each $f \in \mathcal{A}$ there exists (ℓ_j, u_j) such that $\ell_j \le f \le u_j$, *a.e.* with respect to Lebesgue measure and

(ii) $d(\ell_j, u_j) \le \epsilon$, $j = 1, \ldots, m$.

The smallest number of such brackets to cover $\mathcal{A}$ is called the *bracketing number* and is denoted by $N_{[]}(\epsilon, \mathcal{A}, d)$. The *bracketing entropy* is defined by $H_{[]}(\epsilon, \mathcal{A}, d) = \log N_{[]}(\epsilon, \mathcal{A}, d)$.

Generally, if $\mathcal{A}$ is a parametric model of dimension d, then $N_{[]}(\epsilon, \mathcal{A}, d) \sim \epsilon^{-d}$ as can be proved using a Lipschitz argument; see, for example, van der Vaart and Wellner (1996, section 2.7.4). But such arguments require the that derivative of the square root of the density be bounded. This is not the case for mixtures. This is easy to see even in a simple mixture model like $(1-p)\phi(y; 0, 1) + p\phi(y; 1, 2)$; the derivative of the square root of this density is unbounded at $p = 0$. Instead, we must bound the entropy by other methods. The result is given in the following theorem.

THEOREM 1. *Consider the set $\mathcal{M}_k$ defined by (2). If $\epsilon \le 1$, there exists positive constants c_1 and c_2, not depending on k or ϵ, such that*

$$N_{[]}(\epsilon, \mathcal{M}_k, d_H) \le c_1 c_2^k \, m_k^k \left(\frac{S}{s_k}\right)^{2k} \left(\frac{1}{\epsilon}\right)^{3k-1}.$$

To prove Theorem 1, we need some Lemmas.

LEMMA 1. *Let s, m and S be positive constants and define*

$$\mathcal{A} = \{\phi(\cdot\,; \mu, \sigma);\ |\mu| \le m,\ s \le \sigma \le S\}.$$

Then, for $S \ge 1$ and $\epsilon \in (0, 1)$,

$$N_{[]}(\epsilon, \mathcal{A}, d_H) \le \frac{128\,(2m)\left(\frac{S}{s}\right)^2}{\epsilon^2}.$$

PROOF. Let $\delta = \epsilon/2$ and $\tau^2 = (1+\delta)S^2$. Let

$$r = \left\lceil \frac{4\log\left(\frac{S\sqrt{1+\delta}}{s}\right)}{\log(1+\delta)} \right\rceil$$

where $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. Define $\sigma_j^2 = \tau^2(1+\delta)^{-j/2}$ for $j = 2, \ldots, r$. Note that $\sigma_r^2 \le s^2 \le S^2 = \sigma_2^2$. For $j \in \{2, \ldots, r\}$ let $\gamma_j = \delta\sigma_{j-2}/2$, let

$I_j = \lceil m/\gamma_j \rceil$ and let $\mu_{ij} = i\gamma_j$ for $i = -I_j, \ldots, 0, \ldots, I_j$. Note that $[-m, m] \subset [-I_j\gamma_j, I_j\gamma_j]$. For $j \in \{2, \ldots, r\}$ and $-I_j \le i \le I_j$ let

$$B_{ij} = \left\{(\mu, \sigma);\ \mu \in \left[\mu_{ij} - \frac{\delta\sigma_j}{4}, \mu_{ij} + \frac{\delta\sigma_j}{4}\right], \sigma^2 \in \left[\sigma_{j+1}^2, \sigma_j^2\right]\right\}. \tag{4}$$

The B_{ij} cover the parameter space. (A similar construction is used in Tong and Viele (1998).)

Let

$$\ell_{ij}(y) = (1+\delta)^{-1}\phi\left(y;\ \mu_{ij}, \frac{\sigma_{j+1}^2}{(1+\delta)^{1/4}}\right)$$

and

$$u_{ij}(y) = (1+\delta)\phi\left(y;\ \mu_{ij}, \sigma_j^2(1+\delta)\right).$$

We claim that (ℓ_{ij}, u_{ij}) brackets B_{ij}. This follows, after some algebra, from the fact that, whenever $\sigma_1 < \sigma_2$,

$$\frac{\phi(y; \mu_1, \sigma_1)}{\phi(y; \mu_2, \sigma_2)} \le \frac{\sigma_2}{\sigma_1} \exp\left\{\frac{(\mu_1 - \mu_2)^2}{2(\sigma_2^2 - \sigma_1^2)}\right\}.$$

Next we bound $d_H(\ell_{ij}, u_{ij})$. In general, if f and g are probability density functions then $d_H^2((1+\delta)f, (1+\delta)^{-1}g) = d_H^2(f, g) + \delta^2/(1+\delta) \le d_H^2(f, g) + \delta^2$. Also, if $f(y) = \phi(y; \mu_1, \sigma_1)$ and $g(y) = \phi(y; \mu_2, \sigma_2)$ then

$$d_H^2(f, g) = 2\left[1 - \left\{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}\right\}^{1/2} \exp\left\{-\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}\right\}\right].$$

So,

$$\begin{aligned} d_H^2(\ell_{ij}, u_{ij}) &\le 2\left[1 - \sqrt{2}\left\{\frac{(1+\delta)^{7/8}}{(1+\delta)^{7/4} + 1}\right\}^{1/2}\right] + \delta^2 \\ &\le 2\delta^2 \le \epsilon^2. \end{aligned}$$

The last line holds because of the following inequality:

$$1 - \left\{\frac{2u}{1+u^2}\right\}^{1/2} = \frac{(u-1)^2}{\left[(1+u^2)\left(1 + \left\{\frac{2u}{1+u^2}\right\}^{1/2}\right)\right]} \le \frac{1}{2}(u-1)^2$$

where $u = (1+\delta)^{7/8} > 1$ and $u - 1 \le \delta$.

Finally we count the number of boxes N. For each j, the number of boxes is less than or equal to $2m/\gamma_j$. Thus, we see that

$$N \le 2m \sum_{j=2}^{r} \frac{1}{\gamma_j}$$

$$
\begin{aligned}
&= \frac{4m}{\delta S\sqrt{1+\delta}} \sum_{j=2}^{r} (1+\delta)^{(j-2)/4} \\
&= \frac{4m}{\delta S(1+\delta)} \sum_{j=2}^{r} (1+\delta)^{j/4} \\
&\leq \frac{4mr}{\delta S(1+\delta)} (1+\delta)^{r/4} \\
&= \frac{4mr}{\delta S(1+\delta)} \frac{S\sqrt{1+\delta}}{s} \\
&\leq \frac{16m}{\delta s} \frac{\log\left(\frac{S\sqrt{1+\delta}}{s}\right)}{\log(1+\delta)} \leq \frac{32m}{\delta s} \frac{S\sqrt{1+\delta}}{\delta s} \\
&\leq \frac{256mS}{\epsilon^2 s^2} \leq \frac{256mS^2}{\epsilon^2 s^2}. \square
\end{aligned}
$$

Let $\mathcal{S}_{k-1} = \{p = (p_1, \ldots, p_k); p_j \geq 0, \sum_j p_j = 1\}$ be the $k-1$ dimensional simplex.

LEMMA 2. (Bracketing Entropy of the Simplex.) *If* $\epsilon \leq 1$ *then*

$$
N_{[]}(\epsilon, \mathcal{S}_{k-1}, d_H) \leq \frac{\sqrt{2}\sqrt{k+1}(2\pi e)^{k/2}}{\epsilon^{k-1}}.
$$

PROOF. Given $p = (p_1, \ldots, p_k) \in \mathcal{S}_{k-1}$, let $q = (q_1, \ldots, q_k)$ where $q_j = \sqrt{p_j}$. Then $p \in \mathcal{S}_{k-1}$ if and only if $q \in Q^+ \cap U$ where U is the surface of the unit sphere and Q^+ is the positive quadrant of $\mathcal{R}^k$. By virtue of this mapping, an ϵ-L_2 bracketing of $Q^+ \cap U$ corresponds to an ϵ-Hellinger bracketing of $\mathcal{S}_{k-1}$.

Divide the unit cube in $\mathcal{R}^k$ into disjoint cubes with sides parallel to the axes and with sides of length $\epsilon/\sqrt{k}$. Let $\mathcal{C} = \{C_1, \ldots, C_N\}$ be the subset of these cubes that have non-empty intersection with $Q^+ \cap U$. Let b_r be the vertex of C_r furthest from the origin and let a_r be the vertex of C_r closest to the origin. Note that $\sum_j (a_{rj} - b_{rj})^2 = \epsilon^2$ so $\{(a_1, b_1), \ldots, (a_N, b_N)\}$ forms an ϵ-L_2 bracketing. It remains to count the number of cubes N.

Let $T_a = \{q \in Q^+; ||q|| \leq a\}$. Note that $\mathcal{C} \subset T_{1+\epsilon} - T_{1-\epsilon} \equiv A$ so

$$
\text{Volume}(A) \geq \text{Volume}(\mathcal{C}) = N \left(\frac{\epsilon}{\sqrt{k}}\right)^k.
$$

Let $V_k(a) = a^k \pi^{k/2}/\Gamma((k/2)+1)$ denote the volume of a sphere of radius a. Then,

$$
N \leq \frac{\text{Volume}(A)}{\left(\frac{\epsilon}{\sqrt{k}}\right)^k}
$$

$$
\begin{aligned}
&= \frac{1}{2^k}\frac{V_k(1+\epsilon)-V_k(1-\epsilon)}{\left(\frac{\epsilon}{\sqrt{k}}\right)^k} \\
&= \frac{1}{2^k}\frac{(1+\epsilon)^k-(1-\epsilon)^k}{\left(\frac{\epsilon}{\sqrt{k}}\right)^k}\frac{\pi^{k/2}}{\left(\frac{k}{2}\right)!} \\
&\leq \left(\frac{\pi e}{2}\right)^{k/2}\frac{(1+\epsilon)^k-(1-\epsilon)^k}{\epsilon^k}
\end{aligned}
$$

since $x! \geq x^x e^{-x}$.

Now, from the Binomial expansion,

$$
\begin{aligned}
(1+\epsilon)^k-(1-\epsilon)^k &= 2\sum_{j=1}^{\lceil k/2\rceil}\binom{k}{2j-1}\epsilon^{2j-1} \\
&\leq \epsilon\, 2\left\lceil\frac{k}{2}\right\rceil\binom{k}{\lceil\frac{k}{2}\rceil} \\
&= \epsilon\, 2\left\lceil\frac{k}{2}\right\rceil 2^k\prod_{j=1}^{\lceil\frac{k}{2}\rceil}\left(1-\frac{1}{2j}\right) \\
&\leq \epsilon\, 2^k\, 2\left\lceil\frac{k}{2}\right\rceil e^{-\frac{1}{2}\sum_{j=1}^{\lceil\frac{k}{2}\rceil}1/j} \\
&\leq \epsilon\, 2^k\, 2\left\lceil\frac{k}{2}\right\rceil e^{-\frac{1}{2}\log\lceil\frac{k}{2}\rceil} \\
&\leq \epsilon\sqrt{2}2^k\sqrt{k+1}.
\end{aligned}
$$

The conclusion follows. □

LEMMA 3. *Let* $(\ell_1, u_1), \ldots, (\ell_m, u_m)$ *be any* ϵ *Hellinger bracketing. If* $\epsilon \leq 1$ *then*

$$
\Delta_\epsilon \equiv \max_j \int u_j(x)dx \leq 1+6\epsilon.
$$

PROOF. Choose any upper bracket u_j and let ℓ_j be the corresponding lower bracket. Then

$$
\begin{aligned}
\int u_j(x)d\mu(x) &= \int[u_j(x)-\ell_j(x)]d\mu(x)+\int \ell_j(x)d\mu(x) \\
&\leq 2d_{TV}(\ell_j, u_j)+1 \\
&\leq 2\sqrt{\Delta_\epsilon}d_H(\ell_j, u_j)+1 \\
&\leq 2\epsilon\sqrt{\Delta_\epsilon}+1.
\end{aligned}
$$

Take the maximum over all j on both sides of this equation to conclude that $\Delta_\epsilon \leq 2\epsilon\sqrt{\Delta_\epsilon}+1$. This quadratic inequality implies that $\sqrt{\Delta_\epsilon} \leq \epsilon + \sqrt{\epsilon^2+1}$. Square both sides and use the fact that $\sqrt{\epsilon^2+1} \leq 1+\epsilon$. The conclusion follows. □

LEMMA 4. Let $\{\ell_1,\ldots,\ell_k\}$ and $\{u_1,\ldots,u_k\}$ be non-negative, integrable functions and let $(a_1,\ldots,a_k)$ and $(b_1,\ldots,b_k)$ be vectors of non-negative real numbers. Let $\ell = \sum_{j=1}^k a_j\ell_j$ and $u = \sum_{j=1}^k b_j u_j$. Then,

$$d_H^2(\ell,u) \leq \sum_{j=1}^k d_H^2(a_j\ell_j, b_j u_j).$$

PROOF. Note that

$$d_H^2(\ell,u) = \sum_j b_j \int u_j + \sum_j a_j \int \ell_j - 2\int \sqrt{\sum_j b_j u_j \sum_j a_j \ell_j}$$

and

$$\sum_{j=1}^k d_H^2(a_j\ell_j, b_j u_j) = \sum_j b_j \int u_j + \sum_j a_j \int \ell_j - 2\int \sum_j \sqrt{a_j b_j \ell_j u_j}.$$

Thus, it suffices to show that

$$\int \sqrt{\sum_j b_j u_j \sum_j a_j \ell_j} \geq \int \sum_j \sqrt{a_j b_j \ell_j u_j}$$

and for this, it suffices to show that

$$\sqrt{\sum_j b_j u_j \sum_j a_j \ell_j} \geq \sum_j \sqrt{a_j b_j \ell_j u_j} \tag{5}$$

for all x. This follows from the Cauchy-Schwartz inequality. □

THEOREM 2. Let $\mathcal{F}_j = \{f_{\theta_j};\theta_j \in \Omega_j\}$ be a set of density functions for $j=1,\ldots,k$. Let

$$\mathcal{M}_k = \left\{ f(\cdot) = \sum_{j=1}^k p_j f_{\theta_j}(\cdot);\ \theta_j \in \Omega_j, p_j \geq 0, \sum_j p_j = 1 \right\}.$$

If $\epsilon \leq 1$ then

$$N_{[]}(\epsilon, \mathcal{M}_k, d_H) \leq \frac{\sqrt{2(k+1)}(2\pi e)^{k/2}}{(\epsilon/3)^{k-1}} \left(\frac{3}{\epsilon}\right)^{k-1} \prod_{j=1}^k N_{[]}(\epsilon/3, \mathcal{F}_j, d_H). \tag{6}$$

PROOF. Let $\delta = \epsilon/3$ and let

$$\mathcal{B}_j = \{(\ell_{j1}, u_{j1}),\ldots,(\ell_{jm}, u_{jm})\}$$

be a set of δ Hellinger brackets for $\mathcal{F}_j$. Let let $\{(a_1, b_1), \ldots, (a_s, b_s)\}$ be an δ bracketing for the simplex $\mathcal{S}_{k-1}$. Note that each a_r and b_r are vectors of length k. From Lemma 3, $\max_j \int u_j(x)dx$ and $\max_r \sum_{j=1}^k b_{rj}$ are bounded above by $1+6\delta$, where b_{rj} is the j^{th} component of the vector b_r.

Consider $h(x) = \sum_j p_j f_{\theta_j}(x) \in \mathcal{M}_k$. Let $a = (a_1, \ldots, a_k)$ and $b = (b_1, \ldots, b_k)$ be an ϵ-bracket for $p = (p_1, \ldots, p_k)$. Let $(\ell_j, u_j) \in \mathcal{B}_j$ be an ϵ-bracket for f_{θ_j}. Define $\ell = \sum_j a_j \ell_j$ and $u = \sum_j b_j u_j$. Clearly, $\ell(x) \leq h(x) \leq u(x)$ and the number of such brackets is bounded by the right hand side of (6). Now we show that $d_H(\ell, u) \leq \epsilon$.

By Lemma 4, $d_H^2(\ell, u) \leq \sum_{j=1}^k d_j^2$ where $d_j = d_H(b_j u_j, a_j \ell_j)$. Now, using the Cauchy-Schwartz inequality and the fact that $\int \ell_j \leq 1$ we have

$$\begin{aligned}
d_j^2 &= \int \left(\sqrt{a_j \ell_j} - \sqrt{b_j u_j}\right)^2 dx \\
&= \int \left(\sqrt{b_j}\left(\sqrt{\ell_j} - \sqrt{u_j}\right) + \left(\sqrt{a_j} - \sqrt{b_j}\right)\sqrt{\ell_j}\right)^2 dx \\
&= b_j \int \left(\sqrt{\ell_j} - \sqrt{u_j}\right)^2 dx + \left(\sqrt{a_j} - \sqrt{b_j}\right)^2 \int \ell_j dx + 2\sqrt{b_j}\left(\sqrt{b_j} - \sqrt{a_j}\right) \int \sqrt{\ell_j}\left(\sqrt{u_j} - \sqrt{\ell_j}\right) \\
&\leq b_j \delta^2 + \left(\sqrt{a_j} - \sqrt{b_j}\right)^2 + 2\sqrt{b_j}\left(\sqrt{b_j} - \sqrt{a_j}\right)\delta.
\end{aligned}$$

Thus,

$$\begin{aligned}
d_H^2(\ell, u) &\leq \sum_j d_j^2 \\
&\leq \sum_j b_j \delta^2 + \sum_j \left(\sqrt{a_j} - \sqrt{b_j}\right)^2 + 2\delta \sum_j \sqrt{b_j}\left(\sqrt{b_j} - \sqrt{a_j}\right) \\
&\leq \delta^2(1 + 6\delta) + \delta^2 + 2\delta\sqrt{(1 + 6\delta)}\delta \\
&\leq \delta^2[4 + 2\sqrt{3}] \leq \epsilon^2. \square
\end{aligned}$$

The proof of Theorem 1 follows from Lemma 1, Lemma 2 and Theorem 2.

2.2. Large Deviation Bound Now we use the results of the previous section to bound the likelihood ratio. We will need the following result from Wong and Shen (1995).

LEMMA 5. (Wong and Shen, 1995, Theorem 1). *Let $X_1, \ldots, X_n$ be i.i.d. from a distribution P_0 with density f_0 and define the likelihood ratio*

$$R_n(f) = \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}.$$

Let $\mathcal{A}$ be a set of density functions. There are positive constants c_1, c_2, c_3, c_4 such that if

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} H_{[]}^{1/2}(u/c_3, \mathcal{A}, d_H)du \leq c_4\sqrt{n}\epsilon^2$$

then

$$P_0^*\left(\sup_{f\in B} R_n(f) > e^{-nc_1\epsilon^2}\right) \leq 4e^{-c_2 n\epsilon^2}$$

where

$$B = \{f \in \mathcal{A}; d_H(f_0, f) > \epsilon\}.$$

THEOREM 3. *Let $\mathcal{M}_{k_n}$ be a mixture of k_n Gaussians. Let $X_1, \ldots, X_n$ be i.i.d. from a distribution with some density f_0. Let α, α_0, β and β_0 be nonnegative constants such that $2\alpha + \beta \leq 1$ and $2\alpha_0 + \beta_0 \geq 1$. Let $k_n = n^\beta/(\log n)^{\beta_0}$ and $\epsilon_n = (\log n)^{\alpha_0}/n^\alpha$. Suppose that $\frac{m_k}{s_k} = O(k^\eta)$ for some $\eta > 0$. Then, with probability one, there exists n_0, c_1 and c_2 such that, for all $n \geq n_0$,*

$$\sup_{f\in B_n} R_n(f) < c_1 e^{-c_2 n\epsilon_n^2}$$

where $B_n = \{f \in \mathcal{M}_{k_n}; d_H(f_0, f) > \epsilon_n\}$.

PROOF. It follows from Theorem 1 that $N_{[]}(\epsilon_n, \mathcal{F}_k, d_H) \preceq \left(\frac{r}{\epsilon_n}\right)^{3k_n}$ where $r \asymp \left(\frac{S^2 m_{k_n}}{s_{k_n}^2}\right)^{1/3}$. So

$$\begin{aligned} J(\epsilon) \equiv \int_{\epsilon_n^2/2^8}^{\sqrt{2}\epsilon_n} H_{[]}^{1/2}(u/c_3, \mathcal{M}_{k_n}, d_H)du &\preceq \sqrt{k_n}\int_{\epsilon_n^2/2^8}^{\sqrt{2}\epsilon_n}\sqrt{\log\frac{r}{x}}dx \\ &\preceq \epsilon_n\left\{k_n \log\frac{r}{\sqrt{2}\,\epsilon_n}\right\}^{1/2} \end{aligned}$$

With k_n, ϵ_n and m_k/s_k chosen as in the statement of the Theorem, it follows that $J(\epsilon_n) \preceq \sqrt{n}\epsilon_n^2$. The result follows from Lemma 5 and the first Borel-Cantelli Lemma. □

3. SATURATION RATE

In this section we establish the saturation rate of the sieve in both Kullback-Leibler distance and χ^2 distance. It turns out that the saturation rate depends on tail conditions on f_0, and mixtures turn out to require some special treatment. The usual saturation rate arguments used in function estimation theory do not readily apply.

Recall that the true density is assumed to be of the form

$$f_0(y) = \int_0^\infty \int_{-\infty}^\infty \phi(y; \mu, \sigma)\, dP(\mu, \sigma).$$

Let m_k, s_k be sequences of positive real numbers with $s_k > 0$. Define

$$R_k = \{(\mu, \sigma); |\mu| < m_k,\ s_k < \sigma < S\}$$

and let $\delta_k = P(R_k^c)$. The following Lemma will be useful.

LEMMA 6. (Barron-Yang, 1995). *If $f/g \le V$ then*

$$D(f, g) \le [2 + \log V]\ d_H^2(f, g).$$

In what follows, it will be convenient to make a slight change in the definition of the sieve. Specifically, we now define $\mathcal{M}_k$ to be mixtures of $k+1$ components instead of k components. This makes some bookkeeping in Theorem 4 simpler. Note that Theorem 3 is still true with this change.

THEOREM 4. *Let $m_k \to \infty$ and $s_k \to 0$. Let $r_k = \sqrt{8m_k/(ks_k)}$ and that $r_k = o(1)$ as $k \to \infty$ and that*

$$E_P(\sigma^{-1} \mid \sigma < s_k) < \infty$$

for all k . Let a_k be the smallest real number such that

$$E_P\left(\frac{1}{\sigma} e^{\mu^2/a_k^2} \;\middle|\; R_k^c\right) \le a_k \tag{7}$$

where the expectation is taken to be the essential supremum over R_k if $P(R_k^c) = 0$. Then, for $r_k < 2.5$,

$$\inf_{f \in \mathcal{M}_k} D(f_0, f) \le 2\delta_k \left[2 + \log\left(1 + a_k\sqrt{S^2 + a_k^2}\right)\right] + \log(1 + \delta_k) + 2r_k \equiv \omega_k. \tag{8}$$

PROOF. Let

$$f_k(y) = \int_{s_k}^{S} \int_{-m_k}^{m_k} \phi(y; \mu, \sigma)\, dP(\mu, \sigma) + \delta_k \phi_0(y)$$

where $\phi_0(y)$ is a normal density with mean 0 and variance $S^2 + a_k^2$. Given a set of points $\{(\mu_1, \sigma_1), \ldots, (\mu_k, \sigma_k)\}$ and a partition $\{A_1, \ldots, A_k\}$ of R_k, to be chosen later, define

$$g_k = \sum_{j=1}^{k} p_j \phi(y; \mu_j, \sigma_j) + \delta_k \phi_0$$

where $p_j = \int_{A_j} dP$. Let $h_k = g_k/(1+\delta_k)$. Then $h_k \in \mathcal{M}_k$, $D(f_0, h_k) = D(f_0, g_k) + \log(1+\delta_k)$ and

$$D(f_0, g_k) = D(f_0, f_k) + \int f_0 \log \frac{f_k}{g_k}. \tag{9}$$

To bound $D(f_0, f_k)$, it is helpful to first bound f_0/f_k. To this end, let

$$\gamma = \frac{\frac{\mu}{\sigma^2}}{\frac{1}{\sigma^2} - \frac{1}{S^2+a_k^2}}$$

and note that

$$\begin{aligned}
\frac{f_0}{f_k} &= \frac{\int_{R_k} \phi\, dP}{\int_{R_k} \phi\, dP + \delta_k \phi_0} + \frac{\int_{R_k^c} \phi\, dP}{\int_{R_k} \phi\, dP + \delta_k \phi_0} \\
&\le 1 + \frac{1}{\delta_k} \int_{R_k^c} \frac{\phi}{\phi_0}\, dP \\
&\le 1 + \frac{\sqrt{S^2+a_k^2}}{\delta_k} \int_{R_k^c} \frac{1}{\sigma} \exp\left\{ -\frac{1}{2} \left[\frac{(x-\mu)^2}{\sigma^2} - \frac{x^2}{S^2+a_k^2} \right] \right\} dP \\
&\le 1 + \frac{\sqrt{S^2+a_k^2}}{\delta_k} \int_{R_k^c} \frac{1}{\sigma} \exp\left\{ -\frac{1}{2} \left[\frac{1}{\sigma^2} - \frac{1}{S^2+a_k^2} \right] (x-\gamma)^2 \right\} \\
&\qquad \times \exp\left\{ \frac{\mu^2}{\sigma^2 (S^2+a_k^2) \left(\frac{1}{\sigma^2} - \frac{1}{S^2+a_k^2} \right)} \right\} dP \\
&\le 1 + \frac{\sqrt{S^2+a_k^2}}{\delta_k} \int_{R_k^c} \frac{1}{\sigma} \exp\left\{ \frac{\mu^2}{S^2+a_k^2-\sigma^2} \right\} dP \\
&\le 1 + \frac{\sqrt{S^2+a_k^2}}{\delta_k} \int_{R_k^c} \frac{1}{\sigma} e^{\mu^2/a_k^2}\, dP \\
&\le 1 + \sqrt{S^2+a_k^2}\; a_k
\end{aligned}$$

where the last inequality follows from (7). By Lemma 6, and the fact that $d_h^2(f_0, f_k) \le 2 d_{TV}(f_0, f_k) \le 2\delta_k$, we have

$$D(f_0, f_k) \le 2\delta_k \left[2 + \log\left(1 + a_k \sqrt{S^2+a_k^2} \right) \right].$$

Thus we have bounded the first term in (9). Next we bound the second term.

Consider any A_j in the partition. Let $\phi_j(x) = \phi(x;\, \mu_j, \sigma_j)$ and define

$$v^{-2} = \frac{1}{\sigma^2} - \frac{1}{\sigma_j^2} \quad \text{and} \quad \gamma = \frac{\frac{\mu}{\sigma^2} - \frac{\mu_j}{\sigma_j^2}}{\frac{1}{\sigma^2} - \frac{1}{\sigma_j^2}}.$$

Then

$$\int_{A_j} \phi\, dP \;=\; \phi_j \int_{A_j} \frac{\phi}{\phi_j}\, dP$$

$$
\begin{aligned}
&= \phi_j \int_{A_j} \frac{\sigma_j}{\sigma} e^{-\frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2} - \frac{(x-\mu_j)^2}{\sigma_j^2}\right]} dP \\
&= \phi_j \int_{A_j} \frac{\sigma_j}{\sigma} e^{-\frac{1}{2v^2}(x-\gamma)^2} e^{\frac{1}{2}\frac{(\mu-\mu_j)^2}{\sigma_j^2-\sigma^2}} dP \\
&\leq \phi_j \int_{A_j} \frac{\sigma_j}{\sigma} e^{\frac{1}{2}\frac{(\mu-\mu_j)^2}{\sigma_j^2-\sigma^2}} dP.
\end{aligned}
$$

Below, we will show that

$$
\left(\frac{\sigma_j}{\sigma}\right) e^{\frac{1}{2}\frac{(\mu-\mu_j)^2}{\sigma_j^2-\sigma^2}} \leq (1+r_k)^2. \tag{10}
$$

It then follows that

$$
\int_{A_j} \phi \, dP \leq (1+r_k)^2 p_j \phi_j.
$$

Hence,

$$
\begin{aligned}
\frac{f_k}{g_k} &= \frac{\sum_j \int_{A_j} \phi \, dP + r_k \phi_0}{\sum_j p_j \phi_j + r_k \phi_0} \\
&\leq \frac{(1+r_k)^2 \sum_j p_j \phi_j + r_k \phi_0}{\sum_j p_j \phi_j + r_k \phi_0} \\
&\leq (1+r_k)^2
\end{aligned}
$$

and so

$$
\int f_0 \log \frac{f_k}{g_k} \leq 2 \log(r_k + 1) \leq 2 r_k.
$$

It remains to be shown that (10) holds and that $g_k \in \mathcal{M}_k$.

Let

$$
v_1 < v_2 < \cdots < v_J \equiv S.
$$

where

$$
v_j = v_{j-1}\sqrt{1+r_k} \quad \text{and} \quad \sigma_j = v_j \sqrt{1+r_k}.
$$

Here, J is the largest integer such that $s_k \leq v_1$. Also, for $\sigma \in [v_{j-1}, v_j]$, divide $[-m_k, m_k]$ into intervals of length ξ_j where

$$
\xi_j = v_j \sqrt{2 r_k \log(1+r_k)}. \tag{11}
$$

Then, for $\sigma \in [v_{j-1}, v_j]$, $\sigma_j/\sigma \leq \sigma_j/v_{j-1} = (1+r_k)$. Also, $\sigma_j^2 - \sigma^2 \geq r_k v_j^2$. Hence,

$$
\exp\left\{\frac{1}{2}\frac{(\mu-\mu_j)^2}{\sigma_j^2-\sigma^2}\right\} \leq \exp\left\{\frac{1}{2}\frac{\xi_j^2}{r_k v_j^2}\right\} \leq (1+r_k)
$$

using (11). Thus,

$$\frac{\sigma_j}{\sigma}\exp\left\{\frac{1}{2}\frac{(\mu-\mu_j)^2}{\sigma_j^2-\sigma^2}\right\}\le(1+r_k)^2$$

which confirms (10).

To ensure that $g_k \in \mathcal{M}_k$ we have to make sure that the above scheme partitions R_k into no more than k pieces. For fixed σ_j, the number of divisions of μ is

$$\frac{\text{total length}}{\xi_j}=\frac{2m_k}{v_j\sqrt{2r_k\log(1+r_k)}}.$$

So, using the fact that $r_k < 2.5$, the total number N of rectangles is

$$\begin{aligned}
N &= \frac{2m_k}{\sqrt{2r_k\log(1+r_k)}}\left[\frac{1}{v_1}+\cdots\frac{1}{v_J}\right]\\
&= \frac{2m_k}{\sqrt{2r_k\log(1+r_k)}}\left[\frac{1}{s_k}+\frac{1}{s_k(1+r_k)^{1/2}}+\frac{1}{s_k(1+r_k)}+\cdots+\frac{1}{s_k(1+r_k)^{J/2}}\right]\\
&\le \frac{2m_k}{s_k\sqrt{2r_k\log(1+r_k)}}\sum_{j=0}^{\infty}\left(\frac{1}{1+r_k}\right)^{j/2}\\
&= \frac{2m_k}{s_k\sqrt{2r_k\log(1+r_k)}}\frac{\sqrt{1+r_k}}{\sqrt{1+r_k}-1}\\
&\le \left(\frac{m_k}{s_k}\right)\frac{4\sqrt{2}}{\sqrt{r_k\log(1+r_k)}}\frac{1}{r_k}\\
&\le \frac{8m_k}{s_kr_k^2}\le k.\square
\end{aligned}$$

COROLLARY 1. *Let h_k be as defined in the proof of Theorem 3. Then*

$$\int f_0(y)\left(\log\frac{f_0(y)}{h_k(y)}\right)^2dy\le\omega_k\log V_k$$

where

$$V_k=\left(1+a_k\sqrt{S^2+a_k^2}\right)(1+r_k)^2(1+\delta_k).$$

As noted in Wong and Shen (1995), it is sometimes useful to have the saturation rate in a distance that is stronger than Kullback-Leibler. Corollary 2 records the χ^2 saturation rate. Recall that the χ^2 distance is defined by

$$\chi^2(f,g)=\int\frac{(f-g)^2}{g}.$$

LEMMA 7. *If $f/g < V$ then*

$$\chi^2(f,g) \leq 2(1+V^{1/2})^2 d_{TV}(f,g).$$

PROOF.

$$\begin{aligned}
\chi^2(f,g) &= \int \frac{(f-g)^2}{g} \\
&= \int (\sqrt{f}-\sqrt{g})^2 \left(1+\sqrt{\frac{f}{g}}\right)^2 \\
&\leq (1+V^{1/2})^2 d_H^2(f,g) \leq 2(1+V^{1/2})^2 d_{TV}(f,g) \square
\end{aligned}$$

COROLLARY 2. *Under the conditions of Theorem 3,*

$$\rho_k \equiv \inf_{f \in \mathcal{M}_k} \chi^2(f_0,f) \leq 2(1+V_k^{1/2})^2 \left\{ 2\delta_k \left[2+\log(1+a_k\sqrt{S^2+a_k^2}) \right] + 2r_k \right\}$$

where

$$V_k = \left(1+a_k\sqrt{S^2+a_k^2}\right)(1+r_k)^2$$

and a_k is as defined in (7).

PROOF. Define f_k and h_k as in Theorem 4. Then, from the proof of Theorem 4 we see that

$$\frac{f_0}{h_k} = \frac{f_0}{f_k}\frac{f_k}{h_k} \leq V_k.$$

Then, from Lemma 7 and Theorem 4,

$$\begin{aligned}
\rho_k &\leq \chi^2(f_0,h_k) \leq 2(1+V_k^{1/2})^2 d_{TV}(f_0,h_k) \\
&\leq 2(1+V_k^{1/2})^2 D(f_0,h_k) \\
&\leq 2(1+V_k^{1/2})^2 \left\{ 2\delta_k \left[2+\log(1+a_k\sqrt{S^2+a_k^2}) \right] + 2r_k. \right\} \square
\end{aligned}$$

4. RATE OF CONVERGENCE

Here we combine the results of the previous sections to compute the rate.

THEOREM 5. *Let $\alpha, \beta > 0$ and $\alpha_0, \beta_0 \geq 0$ be such that $2\alpha+\beta \leq 1$ and $2\alpha_0+\beta_0 \geq 1$. Let $k_n = n^\beta/(\log n)^{\beta_0}$. Define ω_k as in (8) from Theorem 4 and let*

$$t_n = \max\left\{ \frac{(\log n)^{\alpha_0}}{n^\alpha} \; \frac{2}{c_1^{1/2}}\omega_{k_n}^{1/2} \right\}. \qquad (12)$$

Define r_k as in Theorem 4. Then,

$$P^*(d_H(p_0, \hat{p}_n) > t_n) \leq 4e^{-c_2 n t_n^2} + \frac{4 \log\left((1 + a_{k_n}\sqrt{S^2 + a_{k_n}})(1 + r_{k_n})^2(1 + \delta_k)\right)}{c_1 n t_n^2}. \tag{13}$$

PROOF. This proof parallels the argument in Theorem 3 of Wong and Shen 1995. Let $h_k \in \mathcal{M}_k$ be the density defined in the proof of Theorem 4. Let $B_n = \{p \in \mathcal{M}_{k_n}; d_H(p_0, p) > t_n\}$. Then,

$$\begin{aligned}
P^*(d_H(p_0, \hat{p}_n) > t_n) &\leq P^*\left(\sup_{p \in B_n^c} \prod_{i=1}^n \frac{p(Y_i)}{h_{k_n}(Y_i)} \geq e^{-c_1 n t_n^2/2}\right) \\
&\leq P^*\left(\sup_{p \in B_n^c} \prod_{i=1}^n \frac{p(Y_i)}{p_0(Y_i)} \geq e^{-c_1 n t_n^2}\right) + P\left(\prod_{i=1}^n \frac{p_0(Y_i)}{h_{k_n}(Y_i)} \geq e^{c_1 n t_n^2/2}\right) \\
&= P_1 + P_2.
\end{aligned}$$

Now, $P_1 \leq 4e^{-c_2 n t_n^2}$ by Theorem 3. We bound P_2 using Chebyshev's inequality, Theorem 4 and Corollary 1. Specifically, let $D_n = D(f_0, h_{k_n})$ and let

$$\gamma_n = \int f_0(y) \left(\log \frac{p_0(y)}{h_{k_n}(y)}\right)^2 dy.$$

Note that $D_n \leq \omega_{k_n} \leq (c_1/4)t_n^2$. Define V_k as in Corollary 1. Then,

$$\begin{aligned}
P\left(\prod_{i=1}^n \frac{p_0(Y_i)}{h_{k_n}(Y_i)} \geq e^{c_1 n t_n^2/2}\right) &= P\left(\sum_{i=1}^n \log \frac{p_0(Y_i)}{h_{k_n}(Y_i)} \geq c_1 n t_n^2/2\right) \\
&= P\left(\sum_{i=1}^n \left(\log \frac{p_0(Y_i)}{h_{k_n}(Y_i)} - D_n\right) \geq n[c_1 t_n^2/2 - D_n]\right) \\
&\leq \frac{n \operatorname{Var}\left(\log \frac{p_0(Y)}{h_{k_n}(Y)}\right)}{n^2(c_1 t_n^2/2 - D_n)^2} \\
&\leq \frac{16}{c_1^2 n} \frac{\gamma_n}{t_n^4} \leq \frac{16}{c_1^2 n} \frac{\omega_{k_n} \log V_{k_n}}{t_n^4} \\
&\leq \frac{4}{c_1} \frac{\log V_{k_n}}{n t_n^2}. \square
\end{aligned}$$

Now we consider some special cases.

4.1. Compact Support. An important special case studied in Roeder (1992) is when the mixing measure P has compact support. Thus, suppose that $P(R) = 1$ where $R = \{(\mu, \sigma); s < \sigma < S, -m < \mu < m\}$ and s, S and m are positive constants. This class is still fully nonparametric but the conditions rule out infinite spikes and constrain the density f_0 to have thin tails.

First suppose that we take $m_k \to \infty$ and $s_k \downarrow 0$ in such a way that $m_k/s_k = k^\eta$ for some $\eta \in (0,1)$. By Theorem 4, $\delta_{k_n} = 0$ for large n so that $\omega_{k_n} \sim r_{k_n}$. Hence,

$$\begin{aligned} t_n &\asymp \max\left\{\frac{(\log n)^{\alpha_0}}{n^\alpha}, r_{k_n}^{1/2}\right\} \\ &\asymp \max\left\{\frac{(\log n)^{\alpha_0}}{n^\alpha}, \frac{(\log n)^{\beta_0(1-\eta)/4}}{n^{\beta(1-\eta)/4}}\right\}. \end{aligned}$$

The expression is minimized by taking $\beta = 1 - 2\alpha$, $\beta_0 = 1 - 2\alpha_0$ and $\alpha = \alpha_0 = (1/2)(1 - \eta)/(3 - \eta)$ giving the rate

$$t_n \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{2}\left(\frac{1-\eta}{3-\eta}\right)}$$

which can be made arbitrarily close to $(\log n/n)^{1/6}$. We suspect that the $\log n$ can be eliminated by replacing the bracketing entropy with local entropy; see the comment after Theorem 1 of Wong and Shen (1995).

Now consider choosing m_k and s_k so that $m_k/s_k = (\log k)^\eta$ for $\eta > 0$. Then, an analysis like that above yields the rate

$$\begin{aligned} t_n &\asymp \max\left\{\frac{(\log n)^{\alpha_0}}{n^\alpha}, \frac{(\log n)^{\beta_0/4}}{n^{\beta/4}}(\beta \log n - \beta_0 \log\log n)^{\eta/4}\right\} \\ &\asymp \left(\frac{\log n}{n}\right)^{1/6} (\log n)^{\eta/6}, \end{aligned}$$

where the best rate is obtained by taking $\alpha = 1/6$, $\beta = 1 - 2\alpha$, $\alpha_0 = (1+\eta)/6$ and $\beta_0 = 1 - 2\alpha_0$. In summary, choosing $k_n \sim n^{2/3}/(\log n)^{1/3}$ yields the rate $(\log n)^{(1+\eta)/6}/n^{1/6}$.

Now we consider data based choices of m_k and s_k. A reasonable restriction on m_{k_n} is $\hat{m}_n = \max|X_i|$ and it is easy to show that eventually $[-m, m] \subset [-\hat{m}_n, \hat{m}_n]$. From Roeder (1992), $\hat{m}_n = O(\sqrt{\log n})$ a.s. Next we estimate s. For this, we let $\hat{s}_n$ be the strongly consistent estimate from Theorem 4.2 of Roeder (1992), denoted by $\hat{h}_n$ in her paper. Her model is slightly different than ours but it can be shown there estimate of s is still consistent in our setting. It follows that $\hat{m}_n/\hat{s}_n = O(\sqrt{\log n})$ almost surely for all large n. This corresponds to the above analysis with $\eta = 1/2$ giving a rate $(\log n)^{1/4}/n^{1/6}$.

REMARK. The rates here are quite different than those obtained from, say, kernel density estimation. This is to be expected since maximum likelihood is very tied to the Kullback-Leibler distance. Indeed, Hall (1987) examines kernel density estimation from the perspective of Kullback-Leibler loss and notes that non-standard rates emerge in that setting too. On the other hand, our results give upper bounds on the rate, so we cannot rule out that improved estimates of the rates might be possible.

4.2. Non-Compact Support. To see how the tails of P affect the rate, consider the simplified case where $s = S = 1$, say, so that $f_0(x) = \int \phi(x;\mu,1)dP(\mu)$. Let P have density p with respect to Lebesgue measure. Suppose that p is such that either (i) $p(\mu) \propto e^{-\lambda\mu}$ or (ii) $p(\mu) \propto |\mu|^{-\lambda}$ for $\lambda > 0$ and $|\mu|$ large. In both these cases, the a_k defined in Theorem 4 is infinite, which precludes us from finding a rate. Indeed, we conjecture that the estimate might even be inconsistent. This is not too surprising. To our knowledge, all sieve-based maximum likelihood estimates assume either a compactly supported density or a thin-tailed density. Similar, problems occur with kernel density estimates under Kullback-Leibler loss (Hall, 1987). A remedy for this problem is discussed in the next section.

5. REDUCING SENSITIVITY TO THE TAIL BEHAVIOR OF *P*.

As we have seen, when the mixing P does not have compact support, the rate of convergence is heavily affected by tail behavior of P. The sensitivity to the tails can be mitigated as follows. Let $\psi_0(x)$ be the density of a t-distribution with 1 degree of freedom and scale parameter $\tilde{S} = 4\sqrt{\pi/2}$.

REMARK: In general, can take $\psi(x;\ \lambda,\mu,\sigma)$ to be the density of a t-distribution with $1/\lambda$ degrees of freedom centered on μ with scale, parameter σ. But this extra freedom does not appear to be needed.

Define a new sieve $\tilde{\mathcal{M}}_{k_n}$ by

$$\tilde{\mathcal{M}}_k = \left\{ f(\cdot) = p_0\psi_0(x) + \sum_{j=1}^{k} p_j\phi(\cdot;\ \mu_j,\sigma_j);\ |\mu_j| \le m_k,\ \text{and}\ s_k \le \sigma_j \le S,\ j = 1,\ldots,k \right\}. \tag{14}$$

THEOREM 6. *Assume that the conditions of Theorem 4 hold except instead of a_k we define $\tilde{a}_k = E(\mu^2/\sigma \mid R_k^c)$. Then,*

$$\inf_{f\in\tilde{\mathcal{M}}_k} D(f_0,f) \le 2\delta_k\ [2 + \log(1+\tilde{a}_k)]\ +\ 2r_k\ \equiv \tilde{\omega}_k. \tag{15}$$

PROOF.

For any μ and $\sigma > 0$,

$$\frac{\phi_{\mu,\sigma}(x)}{\psi_0(x)} = \tilde{S}\sqrt{\frac{\pi}{2}}\,\frac{1}{\sigma}\,e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\,[1 + (x - \mu + \mu)^2/\tilde{S}^2] \tag{16}$$

Note that $e^{-au^2}(u + b)^2$ for $a > 0$ and any b attains a maximum that is $\leq (b + 1/\sqrt{a})^2$. Hence,

$$\begin{aligned} \frac{\phi_{\mu,\sigma}(x)}{\psi_0(x)} &\leq \frac{1}{\sigma}\,2\pi\,[1 + (\mu + \sigma\sqrt{2})^2/\tilde{S}^2] && (17)\\ &\leq \frac{\mu^2}{\sigma} \quad \text{for } |\mu| \geq 4 \text{ and } \sigma \leq 1. && (18)\end{aligned}$$

Now, if we replace ϕ_0 by ψ_0 everywhere it appears in the proof of Theorem 4, we have the following. First,

$$\begin{aligned} \frac{f_0}{f_k} &\leq 1 + \frac{1}{\delta_k}\int_{R_k^c} \frac{\phi}{\psi_0}\,dP && (19)\\ &\leq 1 + \frac{1}{\delta_k}\int_{R_k^c} \frac{\mu^2}{\sigma}\,dP && (20)\\ &= 1 + \tilde{a}_k. && (21)\end{aligned}$$

By Lemma 6, it follows that $D(f_0, f_k) \leq 2\delta_k\,[2 + \log(1 + \tilde{a}_k)]$. This bounds the first term in (9). The argument bounding the second term proceeds exactly as before with ψ_0 in place of ϕ_0. This proves the theorem. □

Note that changing ϕ_0 to ψ_0 in the sieve does not increase the bracketing entropy of the sieve, since ψ_0 is a fixed function. In other words, $\{\psi_0\}$ is a set of functions with bracketing number 1 (for all ϵ) and hence contributes a factor of 1 to the product in (6). We have immediately the following analogue of Theorem 5.

THEOREM 7. *Let $\alpha, \alpha_0, \beta, \beta_0 > 0$ be such that $2\alpha + \beta \leq 1$ and $2\alpha_0 + \beta_0 \geq 1$. Let $k_n = n^\beta/(\log n)^{\beta_0}$. Define $\tilde{\omega}_k$ as in (15) from Theorem 4 and let*

$$t_n = \max\left\{\frac{(\log n)^{\alpha_0}}{n^\alpha}, \frac{2}{c_1^{1/2}}\tilde{\omega}_{k_n}^{1/2}\right\}. \tag{22}$$

Define r_k as in Theorem 4. Then,

$$P^*(d_H(p_0, \hat{p}_n) > t_n) \leq 4e^{-c_2 n t_n^2} + \frac{4\,\log\left((1 + \tilde{a}_{k_n})(1 + r_{k_n})^2\right)}{c_1 n t_n^2}. \tag{23}$$

PROOF. Substitute $\tilde{V}_k = (1+\tilde{a}_k)(1+r_k)^2$ for V_k in the proof of Theorem 5 and use the g_k as modified in Theorem 6. The calculation then proceeds exactly as before. □

5.1. Non-compact Support Case Revisited. The revised sieve allows us to compute a rate when P has non-compact support. Consider again the simplified case where $s = S = 1$, so that $f_0(x) = \int \phi(x;\mu,1)dP(\mu)$. Let P have density p with respect to Lebesgue measure such that p has regularly varying tails, i.e., either (i) $p(\mu) \propto e^{-\lambda|\mu|}$ or (ii) $p(\mu) \propto |\mu|^{-\lambda}$, for $\lambda > 0$ and $|\mu|$ large.

In case (i),

$$\delta_k \propto e^{-\lambda m_k} \quad \text{and} \quad \tilde{a}_k \propto [1 + (\lambda m_k + 1)^2], \tag{24}$$

for any $\lambda > 0$. It follows that

$$\begin{aligned} \tilde{\omega}_k &= ce^{-\lambda m_k}\left\{2 + \log(1 + [1 + c'(\lambda m_k + 1)^2])\right\} + r_k, & (25)\\ &\sim c''e^{-\lambda m_k}\log(m_k) + r_k. & (26)\end{aligned}$$

for some constants $c, c', c'' > 0$ and where the second statement holds for large enough m_k. If $m_k = k^\eta$ for $0 < \eta < 1$, then

$$\frac{e^{-\lambda m_k}\log(m_k)}{r_k} \propto \exp[\frac{1}{2}(1-\eta)\ln k - \lambda k^\eta]\log(k) \to 0 \tag{27}$$

as $k \to \infty$. Hence, $\tilde{\omega}_k \asymp r_k$ and we recover the rates from the compact case. On the other hand, if $m_k = \log(k)$, then the size of λ determines the dominant term. Specifically,

$$\frac{e^{-\lambda m_k}\log(m_k)}{r_k} \propto k^{1/2-\lambda}\frac{\log\log k}{\sqrt{\log k}}. \tag{28}$$

If $\lambda \geq 1/2$, r_k again dominates and we recover the rate from the compact case. If on the other hand $\lambda < 1/2$, then $\tilde{\omega}_k \propto k^{-\lambda}\log\log k$. Hence,

$$\tilde{\omega}_{k_n}^{1/2} \propto n^{-\beta\lambda/2}(\log n)^{\beta_0\lambda/2}\sqrt{\log(\beta\log n - \beta_0\log\log n)}. \tag{29}$$

Choosing exponents to calibrate equation (22), gives us that

$$t_n \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{2}\frac{\lambda}{1+\lambda}}\sqrt{\log\log n}. \tag{30}$$

Similarly, in case (ii),

$$\delta_k \propto m_k^{1-\lambda} \quad \text{and} \quad \tilde{a}_k \propto m_k^2, \tag{31}$$

for $\lambda > 3$. If $\lambda \le 3$, then $\tilde{a}_k$ is infinite, and we cannot compute a rate. It follows that when $\lambda > 3$,

$$\tilde{\omega}_k \;=\; c m_k^{-(\lambda-1)} \left[2 + \log(1 + c' m_k^2)\right] \;+\; r_k, \tag{32}$$

$$\sim\; c'' m_k^{-(\lambda-1)} \log(m_k) \;+\; r_k, \tag{33}$$

for some constants $c, c', c'' > 0$.

If $m_k = k^\eta$ for $0 < \eta < 1$, then r_k dominates whenever $\lambda > \frac{1}{2}\frac{1+\eta}{\eta}$ and we recover the rates in the compact case. This happens whenever $\eta > 1/5$ because $\lambda > 3$. If $\lambda \le \frac{1}{2}\frac{1+\eta}{\eta}$, then

$$\tilde{\omega}_{k_n} \;\asymp\; k_n^{-\eta(\lambda-1)} \log k_n \tag{34}$$

$$=\; (\log n)^{\beta_0 \eta(\lambda-1)}\, n^{-\beta\eta(\lambda-1)}\, (\beta \log n - \beta_0 \log\log n). \tag{35}$$

Calibrating exponents as above, the best rate is obtained with $\beta_0 = 0$ and $\beta = 1/[1+\eta(\lambda-1)]$. This yields

$$t_n \asymp \left(\frac{\log n}{n}\right)^{\frac{1}{2}\frac{\eta(\lambda-1)}{1+\eta(\lambda-1)}} (\log n)^{\frac{1}{2}\frac{1}{1+\eta(\lambda-1)}} \;=\; n^{-\frac{1}{2}\frac{\eta(\lambda-1)}{1+\eta(\lambda-1)}} \sqrt{\log n}. \tag{36}$$

Since $\lambda \le \frac{1}{2}\frac{1+\eta}{\eta}$ implies $\eta(\lambda - 1) \le \frac{1}{2}(1-\eta)$ and since $\eta < 1/5$ in this case, the rate is necessary less than $1/6$.

If $m_k = \log k$, then $m_{k_n}^{-(\lambda-1)} = (\beta \log n - \beta_0 \log\log n)^{-(\lambda-1)} \log(\beta \log n - \beta_0 \log\log n)$ while $r_{k_n} \propto (\beta \log n - \beta_0 \log\log n)^{1/2} n^{-1/2}$, so the first term in $\tilde{\omega}_{k_n}$ dominates for large enough k_n. Taking $\beta_0 = 0, \alpha_0 = 1/2$ and any $\alpha, \beta > 0$ satisfying the conditions of the theorem, we obtain

$$t_n \asymp (\log n)^{-(\lambda-1)/2} \sqrt{\log\log n}, \tag{37}$$

where the exponent is negative because $\lambda > 3$.

6. CONCLUDING REMARKS

We have established an upper bound on the rate of convergence for the mixture of Gaussian sieves. From a practical point of view, our results suggest that choosing $k_n \sim n^{2/3}/(\log n)^{1/3}$ components might be a reasonable strategy. Also, our results suggest the importance of including long tailed components in the sieve. The results are also interesting because the entropy calculations and saturation rate are non-standard. We hope that these calculations will be useful for others working in the area of mixture asymptotics.

Finally, we mention three outstanding problems that is the subject of our current work. First, there is the question of whether the rate we have obtained is also a lower bound. Second there is the problem of choosing the number of components k_n from the data. We find the current methods for computing rates when the sieve index is chosen from the data – as in Barron and Yang (1995) for example – do not directly apply to finite mixtures. Third, we believe that some log terms in the rates can be eliminated by using local entropy instead of entropy. Again, for mixtures, calculating the local entropy appears to be non-trivial. We hope to report on these issues in a future paper.

REFERENCES

Banfield, J. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803- 821.

Barron, A. and Yang, Y. (1995). An asymptotic property of model selection criteria. Technical report, Department of Statistics, Yale University.

Chen, J. (1995). Optimal rate of convergence for finite mixture models, *Annals of Statistics,*, 23, 221-233.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 577- 588.

Geman, S. and Hwang, C. (1982), Nonparametric maximum likelihood estimation by the method of sieves, *Annals of Statistics*, 10, 401-414.

Grenander, U. (1981), *Abstract Inference*, Wiley: New York.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *Annals of Statistics*, 15, 1491-1519.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications,* NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5, Institute of Mathematical Statistics, Hayward, CA.

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering.* Marcel-Dekker, New York.

Priebe, C. (1994). Adaptive mixtures. *Journal of the American Statistical Association,* **89**, 796-806.

Robert, C. (1996). Mixtures of distributions: inference and estimation. In: *Markov Chain Monte Carlo in Practice*, (W. Gilks, S. Richardson, D. Spiegelhalter, eds.) Chapman and Hall, London. p. 441- 464.

Roeder, K. (1992). Semiparametric estimation of normal mixture densities. *Annals of Statistics*, 20, 929-943.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association,* 92, 894-902.

Tong, B., and Viele, K. (1998) Mixtures of Normal Linear Regressions. Technical Report, University of Kentucky.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer. New York.

Wong, W. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLE's, *Annals of Statistics*, 23, 339-362.

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213