Nonparametric Estimation of the Mixing Density Using Polynomials

T. Rebafka\textsuperscript{1}\* and F. Roueff\textsuperscript{2}\**

\textsuperscript{1}LPMA, Univ. Paris 6, Paris, France
\textsuperscript{2}Telecom ParisTech, Inst. Telecom, CNRS LTCI, Paris, France

Received April 9, 2015; in final form, July 17, 2015

Abstract—We consider the problem of estimating the mixing density \( f \) from \( n \) i.i.d. observations distributed according to a mixture density with unknown mixing distribution. In contrast to finite mixtures models, here the distribution of the hidden variable is not restricted to a finite set but is spread out over a given interval. We propose an approach to construct an orthogonal series estimator of the mixing density \( f \) involving Legendre polynomials. The construction of the orthonormal sequence varies from one mixture model to another. Minimax upper and lower bounds of the mean integrated squared error are provided which apply in various contexts. In the specific case of exponential mixtures, it is shown that the estimator is adaptive over a collection of specific smoothness classes, more precisely, there exists a constant \( A > 0 \) such that, when the order \( m \) of the projection estimator verifies \( m \sim A \log(n) \), the estimator achieves the minimax rate over this collection. Other cases are investigated such as Gamma shape mixtures and scale mixtures of compactly supported densities including Beta mixtures. Finally, a consistent estimator of the support of the mixing density \( f \) is provided.

Keywords: mixing density, nonparametric estimation, exponential mixture, scale mixture, Gamma shape mixture, polynomial approximation, support estimation.

2000 Mathematics Subject Classification: primary 62G07; secondary 62G20.

DOI: 10.3103/S1066530715030023

1. MIXTURE DISTRIBUTIONS

We consider mixture distributions of densities belonging to some parametric collection \( \{\pi_\theta, \theta \in \Theta\} \) of densities with respect to the dominating measure \( \zeta \) on the observation space \((X, \mathcal{X})\). A general representation of a mixture density uses the so-called mixing distribution and is of the form

\[
\pi_f(x) = \int_{\Theta} f(t) \pi_\theta(x) \mu(d\theta),
\]

where the mixing density \( f \) is a density with respect to some measure \( \mu \) defined on \( \Theta \). To avoid confusion in the notation of \( \pi_\theta \) and \( \pi_f \), we use Greek letters in the indices to denote scalars and latin letters for functions. If \( \mu \) is a counting measure with a finite number of support points \( \theta_k \), then obviously, \( \pi_f \) is a finite mixture distribution of the form \( \sum_{k=1}^{K} p_k \pi_{\theta_k} \). However, if \( \mu \) denotes the Lebesgue measure on \( \Theta \), and if \( \Theta \) is a given interval, say \( \Theta = [a, b] \), then the distribution of the latent variable \( t \) is spread out over this interval and \( \pi_f \) represents a continuous mixture. In this paper we consider continuous mixtures and the problem of identifying the mixing density \( f \) when a sample of the continuous mixture \( \pi_f \) is observed. Note that when \( \theta \) is a location parameter, the problem of estimating \( f \) is referred to as a deconvolution problem, which has received considerable attention in the nonparametric statistics literature since [9].

Continuous mixtures have been used in very numerous and various fields of application. We just give some recent examples to show that continuous mixtures are still of much interest from an application

\*E-mail: tabea.rebafka@upmc.fr
\**E-mail: francois.roueff@telecom-paristech.fr
point of view. The video-on-demand traffic can be modeled by a continuous Poisson mixture for the purpose of efficient cache managing [22]. In time-resolved fluorescence, where photon lifetimes have exponential distribution and parameters depend on the emitting molecules, typically continuous mixtures of exponential distributions are observed [20, 26]. When \( \theta \) is a scale parameter, the distribution \( \pi_f \) is called a scale mixture. Scale mixtures of uniforms are also related to multiplicative censoring introduced in Vardi [27] and length-biased data. A recent application in nanoscience of the latter are length measurements of carbon nanotubes, where observations are partly censored [18]. Exponential mixtures play a significant role in natural sciences phenomena of discharge or disexcitation as e.g. radioactive decays, the electric discharge of a capacitor or the temperature difference between two objects. Several examples of applications of the exponential mixture model can be found in the references of the seminal paper [16].

Not only for applications, but also from the mathematical point of view, scale mixtures are particularly interesting as they define classes of densities that verify some monotonicity constraints. It is well known that any monotone non-increasing density function with support in \((0, +\infty)\) can be written as a mixture of uniform densities \(U[0, \theta][12, \text{ p. 158}]\). Moreover, a \(k\)-monotone density is defined as a non-increasing, convex density function \(h\) with derivatives such that \((-1)^j h^{(j)}\) is nonnegative, non-increasing and convex for all \(j = 1, \ldots, k - 2\). One can show that any \(k\)-monotone density can be represented by a scale mixture of Beta distributions \(B(1, k)\). Furthermore, densities that are \(k\)-monotone for any \(k \geq 1\), also called completely monotone functions, can be written as continuous mixtures of exponential distributions [3].

The literature provides various approaches to the estimation of the mixing density, as for example the nonparametric maximum likelihood estimate (NPMLE). A characteristic feature of this estimator is that it yields a discrete mixing distribution [19, 21]. This appears to be unsatisfactory if we have reasons to believe that the mixing density is indeed a smooth function. In this case a functional approach is more appropriate, which relies on smoothness assumptions on the mixing density \(f\). In Zhang [29] kernel estimators are constructed for mixing densities of a location parameter. Goutis [14] proposes an iterative estimation procedure also based on kernel methods. Aghashian et al. [2] show strong uniform consistency of kernel estimators in the specific case of multiplicative censoring. In the same setting, Andersen and Hansen [1] consider the linear operator \(K\) verifying \(\pi_f = K f\) and estimate \(f\) by an SVD reconstruction in the orthonormal basis of eigenfunctions of \(K\). This approach, although optimal in principle, requires a deep and possibly difficult analysis of the operator \(K\). Polynomial estimators in turn apply easily in many settings. For mixtures of discrete distributions, that is when \(\pi_\theta\) are densities with respect to a counting measure on a discrete space, orthogonal series estimators have been developed and studied in Hengartner [15] and Roueff and Ryden [25]. For such mixtures, these estimators turn out to enjoy similar or better rates of convergence than the kernel estimator presented in Zhang [30]. Comte and Genon-Catalot [6] present a projection estimator based on Laguerre functions that has the specific feature that the support of the mixing density \(f\) is not a compact set as usual, but the entire positive real line. Belomestny and Schoemakers [4] extend the class of scale mixtures and derive estimation methods based on the Mellin transform.

In this paper we show that orthogonal series estimators can be provided in a very general fashion to estimate mixing densities with compact supports. In contrast to Andersen and Hansen [1], who consider only the case of scale mixtures of uniforms, our approach applies to a large variety of continuous mixtures as our numerous examples demonstrate. In the exponential mixture case, in particular, we exhibit an adaptive estimator of the mixing density of an exponential mixture.
2. ESTIMATION METHOD

In this section we develop an orthogonal series estimator and we provide several examples, namely for mixtures of exponential, Gamma, Beta and uniform densities.

2.1. Orthogonal Series Estimator

Throughout this paper the following assumption will be used.

Assumption 1. Let \( \zeta \) be a dominating measure on the observation space \((X, \mathcal{X})\). Let \{\( \pi_\theta, \theta \in \Theta \)\} be a parametric collection of densities with respect to \( \zeta \). Furthermore, let the parameter space \( \Theta = [a, b] \) be a compact interval with known endpoints \( a < b \) in \( \mathbb{R} \). We denote by \( X, X_1, \ldots, X_n \) an i.i.d. sample from the mixture distribution density \( \pi_f \) defined by (1) with \( \mu \) equal to the Lebesgue measure on \([a, b]\).

The basic assumption of our estimation approach is that the mixing density \( f \) in (1) is square integrable, that is \( f \in L^2[a, b] \). Then, for any complete orthonormal basis \( (\psi_k)_{k \geq 1} \) of the Hilbert space \( \mathbb{H} = L^2[a, b] \), the mixing density \( f \) can be represented by the orthogonal series \( f(\theta) = \sum_{k \geq 1} c_k \psi_k(\theta) \), where the coefficients \( c_k \) correspond to the inner products of \( f \) and \( \psi_k \). If we have estimators \( \hat{c}_{n,k} \) of those coefficients, then an estimator of the mixing density \( f \) is obtained by \( \sum_{k=1}^m \hat{c}_{n,k} \psi_k \).

To construct estimators \( \hat{c}_{n,k} \), we remark that the following relation holds: Let \( \varphi(\theta) = \mathbb{E}_{\pi_\theta}[g(X)] = \int_{x \in \mathcal{X}} g(x) \pi_\theta(x) \zeta(dx) \), \( \theta \in [a, b] \).

\[ \varphi(\theta) = \mathbb{E}_{\pi_\theta}[g(X)] = \int_{x \in \mathcal{X}} g(x) \pi_\theta(x) \zeta(dx), \quad \theta \in [a, b]. \tag{2} \]

Suppose that \( \varphi \) belongs to \( \mathbb{H} \). The mean \( \mathbb{E}_{\pi_\theta}[g(X)] \) can be written as the inner product of \( f \) and \( \varphi \). Namely, by the definition of \( \pi_f \) in (1) and Fubini’s theorem,

\[ \mathbb{E}_{\pi_f}[g(X)] = \int_{x \in \mathcal{X}} g(x) \pi_f(x) \zeta(dx) = \int_a^b f(\theta) \int_{x \in \mathcal{X}} g(x) \pi_\theta(x) \zeta(dx) \, d\theta = \langle f, \varphi \rangle_{\mathbb{H}}. \]

Consequently, by the strong law of large numbers, \( \frac{1}{n} \sum_i g(X_i) \) is a consistent estimator of the inner product \( \langle f, \varphi \rangle_{\mathbb{H}} \) based on an i.i.d. sample \( (X_1, \ldots, X_n) \) from the mixture density \( \pi_f \) defined in (1).

We make the following assumption under which the orthogonal series estimator makes sense.

Assumption 2. Assumption 1 holds and there exists a sequence \( (g_k)_{k \geq 1} \) of \( X \rightarrow \mathbb{R} \) functions such that \( (\varphi_k)_{k \geq 1} \) is a dense sequence of linearly independent functions in \( \mathbb{H} \), where \( \varphi_k(\theta) = \mathbb{E}_{\pi_\theta}[g_k(X)] \) as in (2).

We then proceed as follows. Using linear combinations of the \( \varphi_k \)'s, a sequence of orthonormal functions \( \psi_1, \psi_2, \ldots \) in \( \mathbb{H} \) can be constructed, for instance by the Gram–Schmidt procedure. Say that \( \psi_k \) writes as \( \sum_{j=1}^k Q_{k,j} \varphi_j \) with an array \( (Q_{k,j})_{1 \leq j \leq k} \) of real values that are computed beforehand. Then we define estimators of \( c_k = \langle f, \varphi_k \rangle_{\mathbb{H}} = \sum_{j=1}^k Q_{k,j} \varphi_j \) by the empirical means

\[ \hat{c}_{n,k} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k Q_{k,j} g_j(X_i). \]

Finally, for any integer \( m \), an estimator of \( f \) is given by

\[ \hat{f}_{m,n} = \frac{1}{n} \sum_{k=1}^m \hat{c}_{n,k} \psi_k = \frac{1}{n} \sum_{i=1}^n \sum_{j,k,l=1}^m Q_{k,j} Q_{k,l} g_j(X_i) \varphi_l, \tag{3} \]

with the convention \( Q_{k,j} = 0 \) for all \( j > k \). We refer to \( \hat{f}_{m,n} \) as the orthogonal series estimator or the projection estimator of approximation order \( m \).
Define the subspaces
\[ V_m = \text{span} \{ \varphi_1, \ldots, \varphi_m \} \quad \text{for all } m \geq 1. \] (4)

By Assumption 2, the sequence \((V_m)_m\) is strictly increasing, \(V_m\) has dimension \(m\) for all \(m\), and \(\cup_m V_m\) has closure equal to \(H\). By construction the orthogonal series estimator \(\hat{f}_{m,n}\) belongs to \(V_m\). Consequently, the best squared error achievable by \(\hat{f}_{m,n}\) is \(\| f - P_{V_m}(f) \|^2 \), where \(\| \cdot \|_{\sqrt{\delta}}\) denotes the norm associated with \(H\) and \(P_{V_m}\) the orthogonal projection on the space \(V_m\). Hence once the functions \(g_k\) are chosen, the definition of the subspaces \(V_m\) follows and the performance of the estimator will naturally depend on how well \(f\) can be approximated by functions in \(V_m\). It is thus of interest to choose a sequence \((g_k)_{k \geq 1}\) yielding a meaningful sequence of approximation spaces \((V_m)_m\). In the context of scale family mixtures (but not only, see Roueff and Ryden [25]), polynomial spaces appear naturally. Indeed, for any \(m \geq 1\) the orthogonal projection on the space \(H\) holds and there exist two real numbers \(m_1, m_2\) with norms given by
\[ m_1 \leq m \leq m_2 \quad \text{for all } m \text{ such that } \| P_{V_m}(f) \|_{\sqrt{\delta}} \leq m. \]
(6)

For all \(m \geq 1\), we have \(\| P_{V_m}(f) \|_{\sqrt{\delta}} = \| f \|_{\sqrt{\delta}}\), so that, provided that \(\pi_1\) has finite moments, if \(g\) is a polynomial of degree \(k\), so is \(\varphi(\theta) = \pi_1[g(\theta)]\). The following assumption slightly extends this choice for the two following reasons. First, a scale family is not always parameterized by its scale parameter but by its inverse (as for the exponential family). Second, it will appear that the choice of \((g_k)_{k \geq 1}\) not only influences the approximation class (and thus the bias) but also the variance. It may thus be convenient to allow the \(g_k\)'s not to be polynomials, while still remaining in the context of polynomial approximation. This goal is achieved by the following assumption.

**Assumption 3.** Assumption 2 holds and there exist two real numbers \(a' < b'\) and a linear isometry \(T\) from \(H\) to \(H' = L^2[a', b']\) such that, for all \(k \geq 1\), \(T \varphi_k\) is a polynomial of degree \(k - 1\). We denote by \(T^{-1}\) the inverse isometry.

To compute the coefficients \(Q_{k,l}\) under Assumption 3, one may rely on the well-known Legendre polynomials which form an orthogonal sequence of polynomials in \(H' = L^2[a', b']\). Indeed, by choosing \(g_k\) so that \(T \varphi_k\) is the polynomial \(t \mapsto t^{k-1}\), as will be illustrated in all the examples below, the constants \(Q_{k,l}\) are the coefficients of the normalized Legendre polynomials \(\sum_{j=1}^k Q_{k,j} t^{j-1}\). Let us recall the definition of the Legendre polynomials.

**Definition 1** (Legendre polynomials). Let \(a' < b'\) be two real numbers and denote \(\mu = (a' + b')/2\) and \(\delta = (b' - a')/2\). The Legendre polynomials related to the interval \([a', b']\) are defined as the polynomials \(r_k(t) = \sum_{l=1}^k R_{k,l} t^{l-1}\), where the coefficients \(R_{k,l}\) are given by the following recurrence relation
\[ R_{k+1,l} = R_{k,l-1} + \mu R_{k,l} - \beta_k R_{k-1,l} \]
for all \(k, l \geq 1\),
with \(R_{1,1} = 1\) and \(R_{k,l} = 0\) for all \(l > k\), \(\beta_1 = 2\delta\) and \(\beta_k = \delta^2 (k-1)^2 / (4(k-1)^2 - 1)\) for \(k \geq 2\). The obtained sequence \((r_k)_{k \geq 1}\) is orthogonal in \(H' = L^2([a', b'])\) with norms given by \(\| r_k \|_{H'} = \sqrt{\beta_1 \cdots \beta_k}\). Hence, the coefficients of the normalized Legendre polynomials are defined by the relation
\[ Q_{k,l} = \frac{R_{k,l}}{\sqrt{\beta_1 \cdots \beta_k}} \quad \text{for all } k, l \geq 1. \]
(5)

### 2.2. Examples

For illustration we exhibit in this section the orthogonal series estimator in some special cases. Some scale mixtures are presented. As an example for a nonscale mixture we also consider Gamma shape mixtures.

**Example 1 (a). Exponential Mixture.** We first consider continuous exponential mixtures as they play a meaningful role in physics. That is, we consider \(\pi_\theta(x) = \theta e^{-\theta x}\). For the orthogonal series estimator we choose the functions \(g_k(x) = 1 \{ x > k - \frac{1}{2} \} \) for \(k \geq 1\). By (2), we obtain
\[ \varphi_k(\theta) = e^{-(k-\frac{1}{2})\theta}. \]
(6)
We claim that the \( \varphi_k \)'s can be transformed into polynomials in the space \( \mathbb{H}' = L^2[e^{-b}, e^{-a}] \). Indeed, define, for all \( f \in \mathbb{H} = L^2[a, b] \),
\[
Tf(t) = f(-\log t)/\sqrt{t}, \quad t \in [e^{-b}, e^{-a}].
\]
Then one has \( \langle Tf, Tg \rangle_{\mathbb{H}'} = \langle f, g \rangle_{\mathbb{H}} \), hence \( T \) is an isometry from \( \mathbb{H} \) to \( \mathbb{H}' \). Moreover \( T\varphi_k(t) = t^{k-1} \) are polynomials. Denote by \( p_k(t) = \sum_{j=1}^{k} Q_{k,j} t^{j-1} \) the Legendre polynomials in \( \mathbb{H}' \) with coefficients \( Q_{k,j} \) defined by (5) with \( a' = e^{-b} \) and \( b' = e^{-a} \). Denote by \( T^{-1} \) the inverse operator of \( T \) given by \( T^{-1} h(\theta) = e^{-\theta/2} h(e^{-\theta}) \). Since \( T^{-1} \) is a linear isometry, we get that the functions \( \psi_k(t) = T^{-1} p_k(t) = \sum_{j=1}^{k} Q_{k,j} \varphi_j \) are orthonormal in \( \mathbb{H} \). Consequently, an orthonormal series estimator is given by
\[
\hat{f}_{m,n}(\theta) = \frac{1}{n} \sum_{k,j,l=1}^{m} \sum_{i=1}^{n} \left\{ X_i > j - \frac{1}{2} \right\} Q_{k,j} Q_{k,l} e^{-l(t-\frac{1}{2})\theta}. \tag{8}
\]

**Example 1 (b). Exponential Mixture.** The choice of the functions \( g_k \) is not unique and needs to be done with care. For illustration, consider once again exponential mixtures with \( \pi_{\theta}(x) = \theta e^{-\theta x} \). This time we take
\[
g_k(x) = a_k x^k \quad \text{with } a_k = \left( \int x^k \pi_1(dx) \right)^{-1} = 1/k!
\]
and hence \( \varphi_k(\theta) = \theta^{-k} \), for \( k \geq 1 \). To relate \( \varphi_k \) to polynomials, define the isometry \( \tilde{T} \) from \( \mathbb{H} \) to \( \tilde{\mathbb{H}} = L^2[1/b, 1/a] \) by \( \tilde{T} f(t) = \frac{1}{t} f(\frac{1}{t}) \). We have \( \tilde{T} \varphi_k(t) = t^{k-1} \) for all \( k \geq 1 \). Furthermore, denote by \( \tilde{T}^{-1} \) the inverse of \( \tilde{T} \) satisfying \( \tilde{T}^{-1} h(\theta) = \frac{1}{\theta} h(\frac{1}{\theta}) \). Let \( \tilde{p}_k(t) = \sum_{j=1}^{k} Q_{k,j} t^{j-1} \) be the Legendre polynomials in \( \tilde{\mathbb{H}} \) defined with \( a' = 1/b \) and \( b' = 1/a \). Since \( \tilde{T}^{-1} \) is an isometry, \( \psi_k(t) = \tilde{T}^{-1} p_k(t) = \sum_{j=1}^{k} Q_{k,j} \varphi_j(t) \) are orthonormal functions in \( \mathbb{H} \) and the orthonormal series estimator is given by
\[
\hat{f}_{m,n}(\theta) = \frac{1}{n} \sum_{k,j,l=1}^{m} \sum_{i=1}^{n} Q_{k,j} Q_{k,l} \frac{X_i^j}{j!} e^{-l(t-\frac{1}{2})\theta}.
\]

To decide which functions \( g_k \) should be chosen between Example 1 (a) and Example 1 (b), their rates of convergence may be studied. For these examples this will be done later on p. 208.

**Example 2. Gamma Shape Mixture.** Polynomial estimators can be used in the context where the mixed parameter is not necessarily a scale parameter. As pointed out earlier, they have first been used for mixtures on a discrete state space \( X \), such as Poisson mixtures, see [15] and [25]. Let us consider the Gamma shape mixture model. Parametric Gamma shape mixtures have been considered in [28]. For this model \( \pi_{\theta} \) is the Gamma density with shape parameter \( \theta \) and a fixed scale parameter (here set to 1 for simplicity),
\[
\pi_{\theta}(x) = \frac{x^{\theta-1}}{\Gamma(\theta)} e^{-x}, \quad \theta \in [a, b],
\]
where \( \Gamma \) denotes the Gamma function. This model has a continuous state space (\( \zeta \) is the Lebesgue measure on \( \mathbb{R}_+ \)) and is not a scale mixture. We shall construct \( g_k \) and \( \varphi_k(\theta) = E_{\pi_{\theta}}[g_k(X)] \) such that Assumption 3 holds with \( T \) being the identity and \( \varphi_k(\theta) = \theta^{k-1} \). Consider the following sequence of polynomials, \( p_1(t) = 1, p_2(t) = t, \ldots, p_k(t) = t(t+1) \ldots (t+k-2) \) for all \( k \geq 2 \). Since \( p_k(k) \geq 1 \) is a sequence of polynomials of degree \( k - 1 \), there are coefficients \( \tilde{c}_{k,l} \) such that \( t^{k-1} = \sum_{l=0}^{k-1} \tilde{c}_{k,l} t^l \) for \( l = 1, 2, \ldots \). A simple recursive formula for computing \( \tilde{c}_{k,l} \) is provided in Lemma 6 in the Appendix, see Eq. (48). Observe that, for any \( l \geq 1 \),
\[
\int x^{l-1} \pi_{\theta}(x) \, dx = \frac{\Gamma(\theta + l - 1)}{\Gamma(\theta)} = p_l(\theta).
\]
Hence, setting \( g_k(x) = \sum_l c_{k,l} x^{l-1} \), we obtain
\[
\varphi_k(\theta) = \pi_\theta(g_k) = \sum_l \tilde{c}_{k,l} p_l(\theta) = \theta^{k-1},
\]
and thus Assumption 3 holds with \( T \) being the identity operator and \( \varphi_k(\theta) = \theta^{k-1} \). Define \( (Q_{k,l})_{k,l} \) as the coefficients of Legendre polynomials on \( \mathbb{H} = L^2([a, b]) \), that is as in (5) with \( a' = a \) and \( b' = b \). The polynomial estimator defined by (8) reads
\[
\hat{f}_{m,n}(\theta) = \frac{1}{n} \sum_{i=1}^n \sum_{k,j,l=1}^m Q_{k,j} Q_{k,l} \sum_{h=1}^j \tilde{c}_{j,h} X_i^{h-1} \theta^{l-1}.
\]

Example 3. Scale Mixture of Beta Distributions or Uniform Distributions. It is well known that any \( k \)-monotone density, for \( k \geq 1 \), can be represented by a scale mixture of Beta distributions \( B(1, k) \) [3] with
\[
\pi_\theta(x) = \frac{k}{\bar{\theta}} \left( 1 - \frac{x}{\bar{\theta}} \right)^{k-1} \quad \text{for} \quad x \in [0, \theta].
\]
Note that if \( k = 1 \), then \( \pi_\theta \) is the uniform density \( U(0, \theta) \). We take
\[
g_p(x) = a_p x^{p-1} \quad \text{with} \quad a_p = \left( \int x^{p-1} \pi_1(dx) \right)^{-1} = \frac{1}{k \beta(p, k)}, \quad p \geq 1,
\]
where \( \beta(a, b) = \int_0^1 t^{a-1} (1 - t)^{b-1} \, dt \) denotes the Beta function. It follows that \( \varphi_p(\theta) = \theta^{p-1} \). As in the preceding example, if \( f \in \mathbb{H} \) then an orthogonal series estimator \( \hat{f}_{m,n} \) of \( f \) can be constructed by using Legendre polynomials \( p_k(t) = \sum_{j=1}^k Q_{k,j} t^{j-1} \), where the coefficients \( Q_{k,j} \) are defined as in (5) with \( a' = a \) and \( b' = b \). Then according to (3), the corresponding orthogonal series estimator is given by
\[
\hat{f}_{m,n}(\theta) = \frac{1}{n} \sum_{j,p,l=1}^m \sum_{i=1}^n Q_{p,j} Q_{p,l} X_i^{j-1} \frac{1}{k \beta(j, k)} \theta^{l-1}.
\]

In Example 1 (b) we considered the same functions \( g_p \) but here Assumption 3 holds with \( T \) equal to the identity operator on \( \mathbb{H} \). This difference relies on the parametrization of the exponential family by the inverse of the scale parameter.

Example 4. Mixture of exponential distributions with location parameter. The estimator also applies to the deconvolution setting. As an example, consider \( X = Y + \theta \) where \( Y \) and \( \theta \) are independent random variables, \( Y \) has exponential distribution with mean 1 and \( \theta \) has unknown density \( f \) supported on \([a, b]\). The density of \( X \) is given by \( \pi_f(x) = \int_a^b \pi_\theta(x) f(\theta) \, d\theta \) with \( \pi_\theta(x) = e^{-(x-\theta)} \mathbb{1}_{\{x > \theta\}} \). Let \( g_1(x) = 1 \) and
\[
g_k(x) = x^{k-1} - (k - 1)x^{k-2}
\]
for \( k \geq 2 \). Then \( \varphi_k(\theta) = \theta^{k-1} \) for \( k \geq 1 \). The estimator \( \hat{f}_{m,n} \) of \( f \) is then given by
\[
\hat{f}_{m,n}(\theta) = \frac{1}{n} \sum_{i=1}^n \sum_{j,k,l=1}^m Q_{k,l} Q_{k,j} g_j(X_i) \theta^{l-1},
\]
where \( Q_{k,l} \) are the Legendre coefficients defined by (5) with \( a' = a \) and \( b' = b \).

3. ANALYSIS OF THE ORTHOGONAL SERIES ESTIMATOR

In this section the properties of the orthogonal series estimator are analyzed.
3.1. Bias, Variance and MISE

It is useful to write the orthogonal series estimator \( \hat{f}_{m,n} \) defined in (3) in matrix notation. Therefore, we introduce the \( m \times m \) matrix \( Q = (Q_{k,j})_{k,j} \), where \( Q_{k,j} = 0 \) for all \( j > k \), and the \( m \)-vectors

\[
\Phi = [\varphi_1, \ldots, \varphi_m]^T, \quad \Psi = [\psi_1, \ldots, \psi_m]^T = Q \Phi,
\]

\[
g(x) = [g_1(x), \ldots, g_m(x)]^T, \quad \tilde{g} = \frac{1}{n} \sum_{i=1}^{n} g(X_i),
\]

\[
c = [c_1, \ldots, c_m]^T = \langle \Psi, f \rangle_H, \quad \hat{c} = [\hat{c}_{n,1}, \ldots, \hat{c}_{n,m}]^T = Q \tilde{g}.
\]

It follows that the orthogonal series estimator can be written as

\[
\hat{f}_{m,n} = \hat{c}^T \Psi = \tilde{g}^T Q^T Q \Phi.
\]

Further, let \( \Sigma = \pi_f (gg^T) - \pi_f (g) \pi_f (g)^T \) be the covariance matrix of \( g(X_1) \). The MISE is defined by \( E \| \hat{f}_{m,n} - f \|_H^2 \). The orthogonal projection of \( f \) on \( V_m \) is denoted by

\[
P_{V_m} f = c^T \Psi = \sum_{k=1}^{m} c_{n,k} \psi_k.
\]

It is clear that the orthogonal series estimator \( \hat{f}_{m,n} \) is an unbiased estimator of \( P_{V_m} f \). Furthermore, by the usual argument, the MISE is decomposed into two terms representing the integrated variance and integrated squared bias, as summarized in the following result, whose proof is standard and thus omitted.

**Proposition 1.** Suppose that Assumption 2 holds. The orthogonal series estimator \( \hat{f}_{m,n} \) defined in (3) satisfies

(i) For every \( t \in [a, b] \), \( E[\hat{f}_{m,n}(t)] = P_{V_m} f(t) \).

(ii) For every \( t \in [a, b] \), \( \text{Var}(\hat{f}_{m,n}(t)) = \frac{1}{n} \Psi^T(t) Q \Sigma Q^T \Psi(t) \).

(iii) \( E \| \hat{f}_{m,n} - f \|_H^2 = \| P_{V_m} f - f \|_H^2 + \frac{1}{n} \text{tr} (Q \Sigma Q^T) \).

An important issue for orthogonal series estimators \( \hat{f}_{m,n} \) is the choice of the approximation order \( m \). The integrated squared bias \( \| P_{V_m} f - f \|_H^2 \) only depends on how well \( P_{V_m} f \) approximates \( f \), whose rate of convergence depends on the smoothness class to which the density \( f \) belongs. To be more precise, define for any approximation rate index \( \alpha \) and radius \( C \), the approximation class

\[
C(\alpha, C) = \{ f \in H : \| f \|_H \leq C \text{ and } \| P_{V_m} f - f \|_H \leq C m^{-\alpha} \text{ for all } m \geq 1 \}.
\]

So when the mixing density \( f \) belongs to \( C(\alpha, C) \), then the bias of the orthogonal series estimator \( \hat{f}_{m,n} \) is well controlled, namely it decreases at the rate \( m^{-\alpha} \) as \( m \) increases. Furthermore, denote the set of densities in \( H \) by \( H_1 = \{ f \in H : \int_{a}^{b} f(t) \, dt = 1 \} \). We will investigate the rate of convergence of \( \hat{f}_{m,n} \) in \( H \) when \( f \in C(\alpha, C) \cap H_1 \). We will obtain the best achievable rate in the case of exponential mixtures and almost the best one in the case of Gamma shape mixtures.
3.2. Upper Bound of the MISE

We now provide an upper bound of the MISE for the orthogonal series estimator based on Legendre polynomials, that is, when Assumption 3 holds.

To show an upper bound of the MISE we use the following property [25, Lemma A.1]. If \( \lambda > \frac{2 + a' + b'}{b' - a'} + \sqrt{1 + \frac{2 + a' + b'}{b' - a'}} \), then the coefficients of the normalized Legendre polynomials in \( L^2[a', b'] \) defined by (5) verify

\[
\sum_{i=1}^{k} Q_{k,i}^2 = O(\lambda^{2k}) \quad \text{as } k \to \infty. \tag{11}
\]

By combining Proposition 1 (iii) and the bound given in (11) along with a normalization condition on the \( g_k \)'s (Condition (12) or Condition (15) below), we obtain the following asymptotic upper bounds of the MISE.

**Theorem 1.** Let \( \alpha \) be a positive rate index and \( C \) be a positive radius. Suppose that Assumption 3 holds with \( f \in C(\alpha, C) \cap \mathbb{H}_1 \). Let \( \hat{f}_{m,n} \) be defined by (3) with Legendre polynomials coefficients \( Q_{k,j} \) given by (5). Then the two following assertions hold.

(a) Let, for some constants \( C_0 > 0 \) and \( B \geq 1 \),

\[
\Var (g_k(X)) < C_0 B^{2k} \quad \text{for all } k \geq 1. \tag{12}
\]

Set \( m_n = A \log n \) with

\[
A < \frac{1}{2} \left\{ \log B + \log \left( \frac{2 + a' + b'}{b' - a'} + \sqrt{1 + \frac{2 + a' + b'}{b' - a'}} \right) \right\}^{-1}. \tag{13}
\]

Then, as \( n \to \infty \),

\[
\mathbb{E} \left\| \hat{f}_{m,n} - f \right\|_{\mathbb{H}}^2 \leq C^2 m_n^{-2\alpha} (1 + o(1)), \tag{14}
\]

where the o-term only depends on the constants \( \alpha, C, a', b', A \) and \( C_0 \).

(b) Let, for some constants \( C_0 > 0 \) and \( \eta > 0 \),

\[
\Var (g_k(X)) < C_0 k^{\eta k} \quad \text{for all } k \geq 1. \tag{15}
\]

Set \( m_n = A \log n / \log \log n \) with \( A < \eta^{-1} \). Then, as \( n \to \infty \),

\[
\mathbb{E} \left\| \hat{f}_{m,n} - f \right\|_{\mathbb{H}}^2 \leq C^2 m_n^{-2\alpha} (1 + o(1)), \tag{16}
\]

where the o-term only depends on the constants \( \alpha, C, a', b', A \) and \( C_0 \).

**Remark 1.** The larger \( A \), the lower the upper bound in (14). Hence, since \( a', b' \) and \( B \) directly depend on the \( g_k \)'s, the constraint (13) on \( A \) indicates how appropriate the choice of the \( g_k \)'s is.

**Remark 2.** In the examples treated in this paper, \( C_0 \) and \( B \) or \( \eta \) can be chosen independently of \( f \in C(\alpha, C) \cap \mathbb{H}_1 \). Consequently, the bounds given in (14) and (15) show that \( \hat{f}_{m,n} \) achieves the MISE rates \( (\log n)^{-2\alpha} \) and \( (\log(n)/\log \log n)^{-2\alpha} \), respectively, uniformly in \( f \in C(\alpha, C) \cap \mathbb{H}_1 \). In the exponential mixture case, we show below that \( \hat{f}_{m,n} \) of Example 1 (a) is minimax rate adaptive in these classes (since \( m_n \) does not depend on \( \alpha \)). In the Gamma shape mixture case, we could only show that \( \hat{f}_{m,n} \) of Example 2 is minimax rate adaptive in these classes up to the multiplicative \( \log \log n \) term. A data-driven choice of \( m_n \) should not improve the behavior of the estimator, see the case of Poisson mixtures in [6], where a data-driven choice of \( m_n \) is provided.
Proof. We first consider Case (a). By (13), we may choose a number $\lambda$ strictly lying between $\frac{2+\alpha'+\beta'}{\beta'-\alpha'}$ and $e^{1/(2A)}/B$. Note that from Condition (12), it follows by the Cauchy–Schwarz inequality that $|\Sigma_{k,l}| = |\text{Cov}(g_k(X),g_l(X))| \leq C_0 B^l B^l$ for all $k, l$. Thus, we obtain

$$\text{tr} (Q \Sigma Q^T) \leq C_0 \sum_{k=1}^{m} \sum_{l=1}^{k} \sum_{j=1}^{k} |Q_{k,j} Q_{l,j}| B^j B^l \leq C_0 \sum_{k=1}^{m} \left( \sum_{j=1}^{k} Q_{k,j}^2 \sum_{j=1}^{k} B_{l,j}^2 \right) \leq K m^2 \{B \lambda\}^{2m},$$

where the last inequality comes from (11) and $K$ is a positive constant (the multiplicative term $m$ is necessary only for $B = 1$). It follows by the decomposition of the MISE in Proposition 1 (iii) that

$$\mathbb{E}\|\hat{f}_{m,n} - f\|_H^2 \leq C_2 m_n^{-2\alpha} + K n^{-1} m_n(\lambda B)^{2m_n} \leq C_2 m_n^{-2\alpha} \left(1 + \frac{K}{C^2} n^{-1} m_n^{2\alpha+1}(\lambda B)^{2m_n}\right).$$

Now we have for $m_n = A \log n$

$$n^{-1} m_n^{2\alpha+1}(\lambda B)^{2m_n} = A^{2\alpha+1}(\log n)^{2\alpha+1} n^{2A \log B \lambda - 1} = o(1),$$

since $A < 1/(2 \log B \lambda)$.

Let us now consider Case (b). Proceeding as above, for any $\lambda > \frac{2+\alpha'+\beta'}{\beta'-\alpha'} + \sqrt{1 + \frac{2+\alpha'+\beta'}{\beta'-\alpha'}}$, we get

$$\text{tr} (Q \Sigma Q^T) \leq K C_0 \lambda^{2m} m_1^{1+\eta m},$$

which yields

$$\mathbb{E}\|\hat{f}_{m,n} - f\|_H^2 \leq C_2 m_n^{-2\alpha} \left(1 + \frac{K}{C^2} n^{-1} m_n^{2\alpha+1+\eta m \lambda} \lambda^{2m_n}\right).$$

To conclude, it suffices to check that the log of the second term between parentheses tends to $-\infty$ as $n \to \infty$ for $m_n = A \log n/ \log \log n$ with $A < \eta^{-1}$, which is easily done. \[\square\]

Let us check the validity of Condition (12) or Condition (15) for the above examples.

Example 1 (a). Exponential Mixture (continued). Condition (12) immediately holds with $B = C_0 = 1$ for the exponential mixture of Example 1(a) since $g_k(x) = 1\{x > k - \frac{1}{2}\}$.

Example 1 (b). Exponential Mixture (continued). Interestingly, Condition (12) does not hold for Example 1(b), where a different choice of $g_k$’s is proposed. In fact, one finds that $\log \text{Var}(g_k(X))$ is of order $k \log(k)$. Hence, only Condition (15) holds and we fall in case (b) of Theorem 1. Since a slower rate is achieved in this case, this clearly advocates to choose the estimator obtained in Example 1(a) rather than the one in Example 1(b) for the exponential mixture model.

Example 2. Gamma Shape Mixture (continued). We recall that here we set $g_k(x) = \sum_{l=1}^{k} c_{k,l} x^{l-1}$, where the coefficients $(c_{k,l})$ are those defined and computed in Lemma 6 of the Appendix. Using the bound given by (49) in the same lemma, we obtain that $g_k(x) \leq k!(1 \vee |x|^{k-1})$. It follows that

$$\mathbb{E}_{x|\xi}[g_k^2(X)] \leq (k!)^2 (1 + \Gamma(\theta + 2k - 2)/\Gamma(\theta))$$

and $\mathbb{E}_{x|\xi}[g_k^2(X)] \leq (k!)^2 (1 + \Gamma(b + 2k - 2)/\Gamma(b))$ for any $f \in \mathbb{H}_1$. Hence, by Stirling’s formula, we find that Condition (15) holds for $\eta = 4$ and some $C_0$ independent of $f \in \mathbb{H}_1$.

Example 3. Scale Mixture of Beta Distributions or Uniform Distributions (continued). We now verify Condition (12) for Beta mixtures and the $g_p$ of Example 3. Note that we can write $X = AX_0$ with independent random variables $A \sim f$ and $X_0 \sim B(1, k)$. We have for all $p \geq 1$

$$\text{Var}(g_p(X)) \leq \mathbb{E}[X^{2p-2}] = \frac{\mathbb{E}[A^{2p-2}]\mathbb{E}[|X_0|^{2p-2}]}{k^2 \beta^2(p, k)} \leq \frac{b^{2p-2}}{k^2 \beta^2(p, k)} \leq b^{2p-2} k^{2p-2}.$$
Suppose that Assumption of order 0 given by \( \hat{\pi} \) and \( L_1 \) holds with (5) \( \pi \), \( r \), \( \frac{r}{2} \), \( B \), \( C \), \( \in \mathbb{H} \) norm is defined by (17) \( \frac{2}{m} \) with Legendre polynomials coefficients \( Q_{k,j} \) given by (5) achieves the MISE rate \( (\log n)^{-2\alpha} \) uniformly in \( f \in C(\alpha, C) \cap \mathbb{H}_1 \) for any \( \alpha > 0 \) and \( C > 0 \).

Proof. Using the assumptions on \( \pi \) and Jensen’s inequality, we have
\[
B_1^m \leq \int x^m \pi_1(x) \, dx \leq B_2^m \quad \text{for all } m \geq 1,
\]
with \( B_1 = \int x \pi_1(x) \, dx \) and \( B_2 > 0 \) such that the support of \( \pi_1 \) is included in \([0, B_2] \). The result then follows from the same computations as in Example 3.

An immediate consequence of Theorem 1 and Lemma 1 is the following.

**Corollary 1.** Under the assumptions of Lemma 1, the estimator \( \hat{f}_{m,n} \) defined by (3) with Legendre polynomials coefficients \( Q_{k,j} \) given by (5) achieves the MISE rate \( (\log n)^{-2\alpha} \) uniformly in \( f \in C(\alpha, C) \cap \mathbb{H}_1 \) for any \( \alpha > 0 \) and \( C > 0 \).

**Example 4.** Exponential mixture with location parameter (continued). One can show that Condition (15) of Theorem 1 is satisfied, so that the rate of the MISE of the estimator is \( (\log(n) / \log \log n)^{-2\alpha} \). Indeed,
\[
\mathbb{E}[X^r] = r! \int_a^b f(t) \sum_{j=0}^r \frac{t^j}{j!} \, dt \leq r! \sum_{j=0}^r \frac{b^j}{j!} \leq r! e^b,
\]
and thus, using the definition of \( g_k \) in (9), \( \text{Var} (g_k(X)) \leq 2(2k - 2)e^b \approx 2\sqrt{2\pi} e^{b-2k+2(2k-2)^{3/2}} \).

4. APPROXIMATION CLASSES

Although the approximation classes \( C(\alpha, C) \) appear naturally when studying the bias of the orthogonal series estimator defined in (3), it is legitimate to ask whether such classes can be interpreted in a more intuitive way, say using a smoothness criterion. This section provides a positive answer to this question.

4.1. Weighted Moduli of Smoothness

Let us recall the concept of weighted moduli of smoothness as introduced by Ditzian and Totik [8] for studying the rate of polynomial approximations. For \( \alpha < b \in \mathbb{R} \), \( f : [a, b] \to \mathbb{R} \), \( r \in \mathbb{N}^* \) and \( h \in \mathbb{R} \) denote by \( \Delta_h^r(f, \cdot) \) the symmetric difference of \( f \) of order \( r \) with step \( h \), that is
\[
\Delta_h^r(f, x) = \sum_{i=0}^r \binom{r}{i} (-1)^i f(x + (i - r/2)h),
\]
with the convention that \( \Delta_h^0(f, x) = 0 \) if \( x \pm mh/2 \notin [a, b] \). Define the step-weight function \( \varphi \) on the bounded interval \([a, b] \) as \( \varphi(x) = \sqrt{(x-a)(b-x)} \). Then for \( f : [a, b] \to \mathbb{R} \) the weighted modulus of smoothness of \( f \) of order \( r \) and with the step-weight function \( \varphi \) in the \( L^p([a, b]) \) norm is defined as
\[
\omega_r^\varphi(f, t) = \sup_{0 < h \leq t} \| \Delta_h^r(f, \cdot) \|_p.
\]
We recall an equivalence relation of the modulus of smoothness with the so-called $K$-functional, which is defined as

$$K_{r,\varphi}(f,t^r)_p = \inf_{h} \left\{ \|f - h\|_p + t^r \|\varphi^r h^{(r)}\|_p : h^{(r-1)} \in A.C._{loc} \right\},$$

where $h^{(r-1)} \in A.C._{loc}$ means that $h$ is $r-1$ times differentiable and $h^{(r-1)}$ is absolutely continuous on every closed finite interval. If $f \in L^p([a,b])$, then

$$M^{-1}\omega^r_\varphi(f,t)_p \leq K_{r,\varphi}(f,t^r)_p \leq M\omega^r_\varphi(f,t)_p \quad \text{for } t \leq t_0,$$

for some constants $M$ and $t_0$, see Theorem 6.1.1 in Ditzian and Totik [8].

### 4.2. Equivalence Result

We show that the classes $\mathcal{C}(\alpha, C)$ are equivalent to classes defined using weighted moduli of smoothness. This, in turn, will relate them to Sobolev and Hölder classes. To make this precise, we define for constants $\alpha > 0$ and $C > 0$ the following class of functions in $H = L^2([a,b])$

$$\tilde{\mathcal{C}}(\alpha, C) = \{ f \in H : \|f\|_H \leq C \text{ and } \omega^r_\varphi(f,t)_2 \leq Ct^\alpha \text{ for all } t > 0 \},$$

where $\varphi(x) = \sqrt{(x-a)(b-x)}$ and $r = [\alpha] + 1$.

The following theorem states the equivalence of the classes $\mathcal{C}(\alpha, C)$ and $\tilde{\mathcal{C}}(\alpha, C)$. This result is an extension of Proposition 7 in Roueff and Ryden [25] to the case where the subspaces $V_m$ correspond to transformed polynomial classes through an isometry $T$, which includes both a multiplication and a composition with smooth functions.

**Theorem 2.** Let $\alpha > 0$. Suppose that Assumption 3 holds with a linear isometry $T : H = L^2([a,b]) \rightarrow H' = L^2([a',b'])$ given by $Tg = \sigma \times g \circ \tau$, where $\sigma$ is nonnegative and $[\alpha] + 1$ times continuously differentiable, and $\tau$ is $[\alpha] + 1$ times continuously differentiable with a nonvanishing first derivative. Then for any positive number $\alpha$, there exist positive constants $C_1$ and $C_2$ such that for all $C > 0$

$$\mathcal{C}(\alpha, C_1 C) \subset \tilde{\mathcal{C}}(\alpha, C) \subset \mathcal{C}(\alpha, C_2 C),$$

where $\tilde{\mathcal{C}}(\alpha, C)$ is defined in (20) and $\mathcal{C}(\alpha, C')$ is defined in (10) with approximation classes $(V_m)$ given by (4).

For short, we write $\mathcal{C}(\alpha, \cdot) \prec \tilde{\mathcal{C}}(\alpha, \cdot)$ when there exists $C_1 > 0$ such that the first inclusion in (21) holds for all $C > 0$. The validity of both inclusions is denoted by the equivalence $\mathcal{C}(\alpha, \cdot) \asymp \tilde{\mathcal{C}}(\alpha, \cdot)$.

**Proof of Theorem 2.** Weighted moduli of smoothness are used to characterize the rate of polynomial approximation. We start by relating $\mathcal{C}(\alpha, C)$ to classes defined by the rate of polynomial approximation, namely

$$\tilde{\mathcal{C}}(\alpha, C) = \left\{ g \in H' : \|g\|_{H'} \leq C \text{ and } \inf_{p \in \mathcal{P}_m} \|g - p\|_{H'} \leq Cm^{-\alpha}, \text{ for all } m \geq 1 \right\},$$

where $\mathcal{P}_m$ is the set of polynomials of degree at most $m$. Indeed, we see that, since $T$ is a linear isometry,

$$\mathcal{C}(\alpha, C) = \left\{ f \in H : \|f\|_H \leq C \text{ and } \|P_{\mathcal{P}_m} f - f\|_H \leq Cm^{-\alpha} \text{ for all } m \geq 1 \right\} = \left\{ T^{-1}g : g \in H', \|g\|_{H'} \leq C \text{ and } \|P_{\mathcal{P}_m^r} g - g\|_{H'} \leq Cm^{-\alpha} \text{ for all } m \geq 1 \right\} = T^{-1}\tilde{\mathcal{C}}(\alpha, C).$$

As stated in Corollary 7.25 in Ditzian and Totik [8], we have the equivalence $\tilde{\mathcal{C}}(\alpha, \cdot) \asymp \tilde{\mathcal{C}}'(\alpha, \cdot)$, where $\tilde{\mathcal{C}}'(\alpha, C)$ is defined as $\tilde{\mathcal{C}}(\alpha, C)$ but with $a'$ and $b'$ replacing $a$ and $b$. Hence, it only remains to show that

$$T^{-1}\tilde{\mathcal{C}}'(\alpha, \cdot) \asymp \tilde{\mathcal{C}}(\alpha, \cdot).$$
To show this, we use the assumed particular form of $T$, that is $T(g) = \sigma \times g \circ \tau$. Since $T$ is an isometry from $\mathbb{H} = L^2([a, b])$ to $\mathbb{H}' = L^2([a', b'])$ and $\sigma$ is nonnegative, we necessarily have that $\tau$ is a bijection from $[a', b']$ to $[a, b]$ (whose inverse bijection is denoted by $a \circ \tau^{-1}$) and $\sigma = 1/\sqrt{\tau' \circ \tau^{-1}}$. Moreover the inverse isometry writes $T^{-1}(g) = (\sigma \circ \tau^{-1})^{-1} \times g \circ \tau^{-1}$. From the assumptions on $\tau$ we have that $\sigma$, $(\sigma \circ \tau^{-1})^{-1}$, $\tau$ and $\tau^{-1}$ all are $[a] + 1$ times continuously differentiable and the first derivatives of the last two do not vanish. The equivalence (22) then follows by Lemma 5 in the Appendix.

Example 1 (a). Exponential Mixture (continued). In Example 1(a) of continuous exponential mixtures, the operator $T$ is given by (7), that is $\sigma(t) = 1/\sqrt{t}$ and $\tau(t) = -\log t$ and further $\mathbb{H}' = L^2(e^{-b'}, e^{-a'})$. Both $\sigma$ and $\tau$ are infinitely continuously differentiable on $[a, b]$ if $a > 0$, and thus the equivalence given in (21) holds.

Example 1 (b). Exponential Mixture (continued). For the estimator exhibited in Example 1(b) for exponential mixtures, the isometry $T$ is such that $\sigma(t) = \tau(t) = 1/t$ with $a' = 1/b$ and $b' = 1/a$. Hence the conclusion of Theorem 2 holds if $a > 0$.

Examples 2, 3 and 4. Gamma Shape Mixture, Scale Mixture of Beta Distributions and Exponential mixture with location parameter (continued). In the cases of Examples 2, 3 and 4, the transform $T$ is the identity and hence Theorem 2 applies. However, this result is also obtained by Corollary 7.25 in Ditzian and Totik [8].

5. LOWER BOUND OF THE MINIMAX RISK

Our goal in this section is to find a lower bound of the minimax risk

$$\inf_{f \in S_n} \sup_{f \in C} \mathbb{E}_{\pi_f} \| \hat{f} - f \|_{\mathbb{H}},$$

where $S_n$ is the set of all Borel functions from $\mathbb{R}^n$ to $\mathbb{H}$, $C$ denotes a subset of densities in $\mathbb{H}_1$ and $\mathbb{E}_{\pi_f}$ denotes the expectation when $(X_1, \ldots, X_n)$ is an i.i.d. sample with density $\pi_f$ under Assumption 1.

We first provide a general lower bound, which is then used to investigate the minimax rate in the specific cases of exponential mixtures, Gamma shape mixtures and mixtures of compactly supported scale families.

5.1. A General Lower Bound for Mixture Densities

We now present a new lower bound for the minimax risk of mixture density estimation. As in Proposition 2 in [25], it relies on the mixture structure. However, in contrast to this previous result which only applies for mixtures of discrete distributions, we will use the following lower bound in the case of mixtures of exponential distributions, Gamma shape mixtures and scale mixtures of compactly supported densities.

**Theorem 3** (Lower bound). For $f_0 \in \mathbb{H}_1$ and $f_* \in \mathbb{H}$ with $\|f_*\|_{\mathbb{H}} \leq 1$ and $f_0 \pm f_* \in \mathbb{H}_1$ the following lower bound holds, for any $c \in (0, 1)$,

$$\inf_{f \in S_n} \sup_{f \in \{f_0, f_0 \pm f_*\}} \mathbb{E}_{\pi_f} \| f - \hat{f} \|_{\mathbb{H}}^2 \geq c \|f_*\|_{\mathbb{H}}^2 - \frac{c}{(1 - c)^2} \left( \left( 1 + \int |\pi_{f_*}(x)| \, \zeta(dx) \right)^n - 1 \right).$$  \hspace{1cm} (23)

**Proof.** Let $f_*$ be as in the Theorem. We define for a fixed $\hat{f} \in S_n$ and any $c \in (0, 1)$ the set $A = \{ \|f_0 - \hat{f}\|_{\mathbb{H}} \leq \frac{1}{1-c} \}$. Then, for all $\hat{f} \in S_n$, $\sup_{f \in \{f_0, f_0 \pm f_*\}} \mathbb{E}_{\pi_f} \| f - \hat{f} \|_{\mathbb{H}}^2$ is bounded from below by

$$\frac{c}{2} \mathbb{E}_{\pi_{f_0} + f_*} \| f_0 + f_* - \hat{f} \|_{\mathbb{H}}^2 + \frac{c}{2} \mathbb{E}_{\pi_{f_0} - f_*} \| f_0 - f_* - \hat{f} \|_{\mathbb{H}}^2 + (1 - c) \mathbb{E}_{\pi_{f_0}} \| f_0 - \hat{f} \|_{\mathbb{H}}^2$$

$$\geq \frac{c}{2} \mathbb{E}_{\pi_{f_0} + f_*} \left[ \mathbb{1}_{A} \| f_0 + f_* - \hat{f} \|_{\mathbb{H}}^2 \right] + \frac{c}{2} \mathbb{E}_{\pi_{f_0} - f_*} \left[ \mathbb{1}_{A} \| f_0 - f_* - \hat{f} \|_{\mathbb{H}}^2 \right] + (1 - c) \mathbb{E}_{\pi_{f_0}} \| f_0 - \hat{f} \|_{\mathbb{H}}^2.$$
Note that for a function \( k \) defined on \( \mathbb{R}^n \) we have
\[
\mathbb{E}[^{\pi}_{F_0 \pm f_*} k] = \int k(x_1, \ldots, x_n) \prod_{i=1}^{n} (\pi f_0(x_i) \pm \pi f_*(x_i)) \prod_{i=1}^{n} \zeta(dx_i)
\]
\[
= \int k(x_1, \ldots, x_n) \sum_{I,J} [\pm 1]^{\# J} \prod_{j \in J} \pi f_*(x_j) \prod_{i \in I} \pi f_0(x_i) \prod_{i=1}^{n} \zeta(dx_i),
\]
where the sum is taken over all sets \( I \) and \( J \) such that \( I \cup J = \{1, \ldots, n\} \) and \( I \cap J = \emptyset \). Therefore,
\[
\mathbb{E}[^{\pi}_{F_0 + f_*} [1_A \| f_0 + f_* - \hat{f}\|_H^2] + \mathbb{E}[^{\pi}_{F_0 - f_*} [1_A \| f_0 - f_* - \hat{f}\|_H^2]]
\]
\[
= \sum_{I,J} \int \prod_{i \in I} \pi f_0(x_i) \prod_{j \in J} \pi f_*(x_j) [1_A \| f_0 + f_* - \hat{f}\|_H^2 + (-1)^{\# J} \| f_0 - f_* - \hat{f}\|_H^2] \prod_{i=1}^{n} \zeta(dx_i).
\]
Since \( \| f_* \|_H \leq 1 \) and, on \( A \), \( \| f_0 - \hat{f}\|_H \leq \frac{1}{1-c} \), we obtain that, on \( A \), \( \| f_0 \pm f_* - \hat{f}\|_H \leq \| f_0 - \hat{f}\|_H + \| f_* \|_H \leq \frac{1}{1-c} \). This implies that the absolute value of the sum in the last display taken over all sets \( I \) and \( J \) such that the cardinality of \( J \) is positive, \( \# J \geq 1 \), is no greater than
\[
\frac{2}{(1-c)^2} \sum_{I,J: \# J \geq 1} \int \prod_{i \in I} \pi f_0(x_i) \prod_{j \in J} \zeta(dx_j) \prod_{i=1}^{n} \zeta(dx_i)
\]
\[
= \frac{2}{(1-c)^2} \sum_{I,J: \# J \geq 1} \int \prod_{i \in I} \pi f_0(x_i) \prod_{j \in J} \zeta(dx_j) \prod_{i=1}^{n} \zeta(dx_i)
\]
\[
= \frac{2}{(1-c)^2} \left( 1 + \int \left| \pi f_*(x) \| \zeta(dx) \right|^n \right) - 1.
\]
Moreover, the term with \( \# J = 0 \) is 0.
\[
\mathbb{E}[^{\pi}_{F_0} [1_A \| f_0 + f_* - \hat{f}\|_H^2 + \| f_0 - f_* - \hat{f}\|_H^2]] = 2\mathbb{E}[^{\pi}_{F_0} [1_A \| f_0 - \hat{f}\|_H^2 + \| f_* \|_H^2]],
\]
by the Parallelogram law. By combining these results, the minimax risk is bounded from below by
\[
(1-c)\mathbb{E}[^{\pi}_{F_0} \| f_0 - \hat{f}\|_H^2 + c\mathbb{E}[^{\pi}_{F_0} [1_A \| f_0 - \hat{f}\|_H^2 + \| f_* \|_H^2]]
\]
\[
= \frac{c}{(1-c)^2} \left[ 1 + \int \left| \pi f_*(x) \| \zeta(dx) \right|^n \right] - 1.
\]
Finally we see that
\[
(1-c)\| f_0 - \hat{f}\|_H^2 + c1_A \| f_0 - \hat{f}\|_H^2 + \| f_* \|_H^2
\]
\[
= c1_A \| f_* \|_H^2 + ((1-c) + c1_A) \| f_0 - \hat{f}\|_H^2 \geq c1_A \| f_* \|_H^2 + c1_A \| f_* \|_H^2,
\]
where we used \( 1 \geq \| f_* \|_H^2 \). This yields the lower bound asserted in the theorem. \( \square \)

### 5.2. Application to Polynomial Approximation Classes

The lower bound given in (23) relies on the choice of a function \( f_* \) such that \( f_0 \) and \( f_0 \pm f_* \) are in the smoothness class of interest. In this subsection, we give conditions which provide a tractable choice of \( \| f_* \|_H \leq 1 \) for the class \( \mathcal{C}(\alpha, C) \) defined in (10). Following the same lines as Theorem 1 in [25], the key idea consists in restricting our choice using the space \( V_{m \alpha}^2 \) (the orthogonal complement of \( V_m \) in \( H \)) and to control separately the two terms that appear in the right-hand side of (23) within this space.

An important constraint on \( f_* \) is that \( f_0 \pm f_* \in H_1 \). In particular, for controlling the sign of \( f_0 \pm f_* \), we use the following semi-norm on \( H \),
\[
\| f \|_{\infty, f_0} = \text{ess sup}_{t \in \Theta} \frac{|f(t)|}{f_0(t)}.
\]

MATHEMATICAL METHODS OF STATISTICS Vol. 24 No. 3 2015
with the convention $0/0 = 0$ and $s/0 = \infty$ for $s > 0$. Further, for any subspace $V$ of $\mathbb{H}$, we denote
\[ K_{\infty,f_0}(V) = \sup\{||f||_{\infty,f_0} : f \in V, ||f||_{H} = 1\}. \]
The following lemma will serve to optimize the term $||f_s||_H$ on the right-hand side of (23). It is similar to Lemma 2 in [25], so we omit its proof.

**Lemma 2.** Suppose that Assumption 2 holds. Let $f_0$ be in $\mathbb{H}_1$, $\alpha, C_0 > 0$, $K \leq 1$ and let $\mathcal{C}(\alpha, C_0)$ be defined by (10) with $V_m$ given by (4). Let moreover $w \in \mathbb{H}$. Then there exists $g \in \mathcal{C}(\alpha, C_0) \cap V_m^+ \cap w^+$ such that $||g||_{\infty,f_0} \leq K$ and
\[ ||g||_H = \min\left(C_0(m + 1)^{-\alpha}, \frac{K}{K_{\infty,f_0}(V_m^+ \cap V_m^+ \cap w^+)}\right). \]

Under Assumption 3, where the orthonormal functions $\psi_k$ are related to polynomials in some space $\mathbb{H}' = L^2[a', b']$, the constant $K_{\infty,f_0}(V_m^+ \cap V_m^+ \cap w^+)$ can be bounded by $K_{\infty,f_0}(V_{m+2})$ and then using the following lemma.

**Lemma 3.** Suppose that Assumption 3 holds. Let $f_0$ be in $\mathbb{H}_1$ and suppose that
\[ \sup \{||f||_{\infty,f_0} : f \in \mathbb{H} \text{ such that } \sup_{t \in [a', b']} |Tf(t)| \leq 1\} < \infty. \] (24)
Then there exists a constant $C_0 > 0$ satisfying
\[ K_{\infty,f_0}(V_{m+2}) \leq C_0 m \quad \text{for all} \ m \geq 1. \]

**Proof.** Note that $\{Tf : f \in V_m\}$ is the set of polynomials in $\mathbb{H}'$ of degree at most $m - 1$, denoted by $\mathcal{P}_{m-1}$. Using $||f||_H = ||Tf||_{\mathbb{H}'}$ and denoting by $B$ the left-hand side of (24), we have
\[ K_{\infty,f_0}(V_m) = \sup\{||f||_{\infty,f_0} : f \in V_m, ||f||_H = 1\} \]
\[ \leq B \sup\left\{ \sup_{t \in [a', b']} |Tf(t)| : f \in V_m, ||f||_H = 1\right\} \]
\[ = B \sup\left\{ \sup_{t \in [a', b']} |p(t)| : p \in \mathcal{P}_{m-1}, \int_{a'}^{b'} p^2(t) dt = 1\right\}. \]
By the Nikolskii inequality (see, e.g., DeVore and Lorentz [7], Theorem 4.2.6), there exists a constant $C > 0$ such that the last sup is at most $Cm$. Hence there exists $C_0 > 0$ such that $K_{\infty,f_0}(V_m) \leq C_0 m$ for all $m \geq 1$. \hfill $\square$

Theorem 3 and Lemmas 2 and 3 yield the following result.

**Corollary 2.** Let $\alpha \geq 1$ and $C > (b - a)^{-1/2}$. Suppose that Assumption 3 holds with an isometry $T$ satisfying the assumptions of Theorem 2. Let $w$ be an $[\alpha] + 1$ times continuously differentiable function defined on $[a, b]$ and set
\[ v_m = \sup_{g \in V_m^+, ||g||_{H} \leq 1} \int |\pi_w g(x)| \zeta(dx). \] (25)
Then there exists a small enough $C_\ast > 0$ and $C^\ast > 0$ such that, for any sequence $(m_n)$ of integers increasing to $\infty$ satisfying $v_{m_n} \leq C_\ast n^{-1} m_n^{-\alpha}$, we have
\[ \inf_{f \in \mathcal{S}_n} \sup_{f \in \mathcal{E}(\alpha, C) \cap \mathbb{H}_1} \mathbb{E}_{\pi_f} ||\hat{f} - f||^2_{\mathbb{H}} \geq C^\ast m_n^{-2\alpha}(1 + o(1)), \] (26)
where $\mathcal{E}(\alpha, C)$ is the smoothness class defined by (20).
Remark 3. The assumption $C > (b - a)^{-1/2}$ is necessary, otherwise $\hat{C}(\alpha, C) \cap \mathbb{H}_1$ is reduced to one density for $C = (b - a)^{-1/2}$ and is empty for $C < (b - a)^{-1/2}$. To see why, observe that for any $f \in \mathbb{H}_1$, by Jensen's inequality, $\|f\|_H = \int_a^b f^2(t) \, dt \geq (b - a)^{-1}$, with equality implying that $f$ is the uniform density on $[a, b]$.

Proof. We apply Theorem 3 with $f_0$ set as the uniform density on $[a, b]$ and $f_*$ chosen as follows. For some $C_0 > 0$ and an integer $m$ to be determined later, we choose $f_* = w_0$, where $g$ is given by Lemma 2 with $K = \min(1, \sup_{t \in [a, b]} |w(t)|)$. Since $g \in w^+$ and $\|g\|_{\infty, f_0} \leq K$, we get that $f_0 \pm f_* \in \mathbb{H}_1$.

Now we show that $\{f_0, f_0 \pm f_*\} \subset \tilde{C}(\alpha, C)$ for a well-chosen $C_0$. We have $\|f_0\|_H = (b - a)^{-1/2}$ and, since the symmetric differences of all order vanish on $f_0$, we get that $f_0 \in \tilde{C}(\alpha, (b - a)^{-1/2})$. By definition of $g$ in Lemma 2 and Lemma 5 successively, we get that $f_* \in \tilde{C}(\alpha, C_0)$ for some $C_1 > 0$ not depending on $C_0$. Choosing $C_0 = (C - (b - a)^{-1/2})/C_1$, we finally get that

$$\{f_0, f_0 \pm f_*\} \subset \tilde{C}(\alpha, C) \cap \mathbb{H}_1.$$ 

By Lemma 2, $\|g\|_H \to 0$ as $m \to \infty$ and, since $w$ is bounded, it implies that $\|f_*\|_H \leq 1$ for $m$ large enough. Hence we may apply Theorem 3 and, to conclude the proof, it remains to provide a lower bound of the right-hand side of (23) for the above choice of $f_*$. Under the assumptions of Theorem 2, Condition (24) clearly holds. So Lemma 3 and the definition of $g$ in Lemma 2 give that

$$\|g\|_H \leq C_0 m^{-\alpha}$$

for some constant $C_0 > 0$. By definition of $v_m$ and since $g \in V_m^\perp$, we have

$$\int |\pi_{f_0}(x)| \, \zeta(dx) \leq \|g\|_H v_m \leq C_0 m^{-\alpha} v_m.$$

We now apply the lower bound given by (23) with $m = m_n$ for $(m_n)$ satisfying $v_{m_n} \leq C_n n^{-1} m_n^{\alpha}$. We thus obtain

$$\inf \sup \mathbb{E}_x \|\tilde{f} - f\|^2_H \geq c(C_0 m_n^{-\alpha})^2 - \frac{c}{(1-c)^2} C_n C_0 m_n^{-2\alpha} (1 + o(1))$$

$$\geq C^* m_n^{-2\alpha} (1 + o(1)),$$

where the last inequality holds for some $C^* > 0$ provided that $C_*$ is small enough.}

To apply Corollary 2, one needs to investigate the asymptotic behavior of the sequence $(v_m)$ defined in (25). The following lemma can be used to achieve this goal.

Lemma 4. Under Assumption 3, if $\pi_*(x) \in \mathbb{H}$ for all $x \in X$, then $v_m$ defined in (25) satisfies

$$v_m \leq \int \|T[w\pi_*(x)] - P_{\mathcal{P}_{m-1}}(T[w\pi_*(x)])\|_{\mathbb{H}} \, \zeta(dx),$$

(27)

where $\mathcal{P}_{m-1}$ is the set of polynomials of degree at most $m - 1$ in $\mathbb{H}$ and $P_{\mathcal{P}_{m-1}}$ denotes the orthogonal projection in $\mathbb{H}$ onto $\mathcal{P}_{m-1}$.

Proof. Let $g \in V_m^\perp$ such that $\|g\|_H \leq 1$. Then we have, for all $x \in \mathbb{R}$,

$$\pi_{w_0}(x) = \langle w_0, \pi_*(x) \rangle_{\mathbb{H}} = \langle g, \pi_*(x) \rangle_{\mathbb{H}} = \langle Tg, T[w\pi_*(x)] \rangle_{\mathbb{H}}.$$ 

Recall that $TV_m = \mathcal{P}_{m-1}$ is the set of polynomials of degree at most $m - 1$ in $\mathbb{H}$. Hence $Tg$ is orthogonal to $\mathcal{P}_{m-1}$, and for any $p \in \mathcal{P}_{m-1}$, we get, for all $x \in \mathbb{R}$,

$$|\pi_{w_0}(x)| = |\langle Tw, T[w\pi_*(x)] - p \rangle_{\mathbb{H}}| \leq \|T[w\pi_*(x)] - p\|_{\mathbb{H}},$$

(28)

where we used the Cauchy–Schwarz inequality and $\|Tg\|_{\mathbb{H}} = \|g\|_{\mathbb{H}} \leq 1$. Now the bound given by (27) is obtained by taking $p$ equal to the projection of $[w\pi_*(x)]$ onto $\mathcal{P}_{m-1}$ (observe that the right-hand side of (28) is then minimal).
5.3. Minimax Rate for Exponential Mixtures

Consider the exponential case, that is, let Assumption 1 hold with \( \zeta \) defined as the Lebesgue measure on \( \mathbb{R}_+ \), \( \Theta = [a, b] \subset (0, \infty) \) and \( \pi_\theta(x) = \theta e^{-\theta x} \). Let \( C > (b - a)^{-1/2} \) and \( \alpha > 1 \) and define \( \tilde{C}(\alpha, C) \) as in (20). Then there exists \( C^* > 0 \) such that

\[
\inf_{f \in S_n} \sup_{f \in \mathcal{C}(\alpha, C) \cap H_1} \mathbb{E}_{\pi_f} \| \hat{f} - f \|_2^2 \geq C^*(\log n)^{-2\alpha}(1 + o(1)).
\]

**Proof.** Let \( g_k(x) = 1 \{ x > k - \frac{1}{2} \} \) for \( k \geq 1 \). Then Assumption 3 holds with \( \varphi_k \) and \( T \) defined by (6) and (7) respectively. Since \( a > 0 \), \( T \) satisfies the assumptions of Theorem 2. We may thus apply Corollary 2 with \( w = \mathbb{1}_{[a,b]} \). Hence the minimax lower bound given in (29) follows from (26), provided that we have, setting \( m_n = C' \log n \) for some \( C' > 0 \),

\[
v_m = o(n^{-1} m^{-\alpha}) \quad \text{as} \quad n \to \infty,
\]

where \( v_m \) is defined by (25). Note that \( \pi_\theta(x) = \theta e^{-\theta x} \mathbb{1}_{[a,b]}(x) \). We apply Lemma 4 to bound \( v_m \). Using the definition of \( T \) in (7), we have \( [T \pi_\theta(x)](t) = -\log t \), \( t \geq x \), as the sum of its entire and decimal parts, \( x = [x] + \langle x \rangle \), and observe that, since \( x - 1/2 \in [-1/2, 1/2) \) and \( [a', b'] = [e^{-b}, e^{-a}] \subset (0,1) \), the expansion of \( t(x)^{-1/2} = \sum_{k \geq 0} \alpha_k(x)(1-t)^k \) as a power series about \( t = 1 \) satisfies \( \alpha_k(x) = \prod_{j=1}^{k} \langle x - 1/2 - j/k \rangle \leq 1 \). Extending \( -\log t \) about \( t = 1 \), we thus get

\[
-\log(t) t^{-1/2} = \sum_{k \geq 0} \beta_k(x)(1-t)^k
\]

with \( |\beta_k(x)| = |\sum_{j=1}^{k} \alpha_k-t/j| \leq 1 + \log(k) \). For any \( x < m \), we use this expansion to approximate \( [T \pi_\theta(x)](t) = -\log(t) t^{-1/2} \times t[x] \) by a polynomial of degree \( m \). Namely, we obtain

\[
\sup_{t \in [a', b']} \left| [T \pi_\theta(x)](t) - \sum_{k=0}^{m-|x|} \beta_k(x) t^{k+|x|} \right| \leq \sum_{k>m-|x|} (1+ \log(k))(b')^{k+|x|} \leq C_1 c^m,
\]

where we used the bound \( 1 + \log(k) \leq C_1 (c/b')^k \) valid for some constants \( C_1 > 0 \) and \( c \in (b', 1) \) not depending on \( x \). This bound also applies to \( \mathbb{E}_{\pi_\theta} \| [T \pi_\theta(x) - P_{m-1}(T \pi_\theta(x))]|_{\mathbb{H}'} \) by definition of the projection \( P_{m-1} \). For \( x \geq m \), we simply observe that \( \| [T \pi_\theta(x)](t) \|_{\mathbb{H}'} \leq -\log(a') b x^{-1/2} \). This also provides an upper bound for \( \mathbb{E}_{\pi_\theta} \| T \pi_\theta(x) - P_{m-1}(T \pi_\theta(x))|_{\mathbb{H}'} \). Finally, integrating on \( x \geq 0 \) we get

\[
\int_{\mathbb{R}_+} \| [T \pi_\theta(x) - P_{m-1}(T \pi_\theta(x))]|_{\mathbb{H}'} \, dx \leq C_2 m c^m
\]

with constants \( C_2 > 0 \) and \( c < 1 \) not depending on \( m \), and this upper bound applies to \( v_m \) by Lemma 4. This shows that (30) holds provided that \( C' > 0 \) is taken small enough. This completes the proof.

5.4. Minimax Rate for Gamma Shape Mixtures

In this section, we show that in the case of Gamma shape mixtures the orthogonal series estimator of Example 4 achieves the minimax rate up to the log log \( n \) multiplicative term.

**Theorem 5.** Consider the Gamma shape mixture case, that is, let Assumption 1 hold with \( \zeta \) defined as the Lebesgue measure on \( \mathbb{R}_+ \), \( \Theta = [a, b] \subset (0, \infty) \) and \( \pi_\theta(x) = x^{\theta-1} e^{-\theta x}/\Gamma(\theta) \). Let \( C > (b - a)^{-1/2} \) and \( \alpha > 1 \) and define \( \tilde{C}(\alpha, C) \) as in (20). Then there exists \( C^* > 0 \) such that

\[
\inf_{f \in S_n} \sup_{f \in \mathcal{C}(\alpha, C) \cap H_1} \mathbb{E}_{\pi_f} \| \hat{f} - f \|_2^2 \geq C^*(\log n)^{-2\alpha}(1 + o(1)).
\]

\[\text{MATHEMATICAL METHODS OF STATISTICS} \quad \text{Vol. 24 No. 3 2015}\]
Proof. We proceed as in the proof of Theorem 4. This time we set $g_k(x) = \sum_{l=1}^{k} \hat{c}_{k,l} x^{l-1}$ with coefficients $(\hat{c}_{k,l})$ defined in Lemma 6. Assumption 3 then holds with $\mathbb{H}' = \mathbb{H}$ and $T$ defined as the identity operator. Applying Corollary 2 with $w(t) = \Gamma(t)$, we obtain the lower bound given in (31) provided that Condition (30) holds with $m_{n} = C' \log n/\log \log n$ for some $C' > 0$. Again we use Lemma 4 to check this condition in the present case. To this end we must, for each $x > 0$, provide a polynomial approximation of $w(\theta)\pi_{\theta}(x) = x^{\theta-1}e^{-x}$ as a function of $\theta$. Expanding the exponential function as a power series, we get

$$
\sup_{\theta \in [a,b]} \left| w(\theta)\pi_{\theta}(x) - e^{-x} \sum_{k=0}^{m-1} \frac{\log^{k}(x)}{k!} (\theta - 1)^{k} \right| \leq e^{-x} \sum_{k \geq m} \frac{\log^{k}(x)}{k!} e^{k},
$$

where $c = \max(|a-1|, |b-1|)$. Let $(\alpha_m)$ be a sequence of real numbers tending to infinity. The right-hand side of the previous display is less than $e^{c\log(x) - x (c \log(x))/m}$. We use this for bounding $\|w\pi(x) - P_{m-1}(w\pi(x))\|_{\mathbb{H}}$ (recall that $T$ is the identity and $\mathbb{H}' = \mathbb{H}$) when $x \in [e^{-\alpha_m}, \alpha_m]$. When $x \in (0, e^{-\alpha_m})$ we use that the latter is bounded by $O(1)$ and when $x > \alpha_m$ by $O(e^{-x/2})$. Hence Lemma 4 gives that

$$
v_{m} = O(e^{-\alpha_m}) + \frac{e^{m}}{m} \int_{e^{-\alpha_m}}^{\alpha_m} e^{c\log(x) - x} |\log(x)|^{m} \, dx + O(e^{-x/2}).
$$

Now observe that, as $x_{m} \to \infty$, separating the integral $\int_{e^{-\alpha_m}}^{\alpha_m} \, dx$ as $\int_{e^{-\alpha_m}}^{1} \, dx + \int_{1}^{\alpha_m} \, dx$, we get

$$
\int_{e^{-\alpha_m}}^{\alpha_m} e^{c\log(x) - x} |\log(x)|^{m} \, dx = O(e^{\alpha_m} \alpha_{m}^{m}) + O(\log^{m}(\alpha_{m})).
$$

Set $x_{m} = c_{0}m$. By Stirling’s formula, for $c_{0} > 0$ small enough, we get $v_{m} = O(e^{c_{0}m})$ with $c_{1} \in (0, 1)$. We conclude as in the proof of Theorem 4. \hfill \Box

5.5. Lower Bound for Compactly Supported Scale Families

We derived in Corollary 1 an upper bound of the minimax rate for estimating $f$ in $C(\alpha, C)$. It is thus legitimate to investigate whether, as in the exponential mixture case, this upper bound is sharp for mixtures of compactly supported scale families. A direct application of Corollary 2 provides the following lower bound, which, unfortunately, is far from providing a complete and definite answer.

Theorem 6. Consider the case of scale mixtures of a compactly supported density on $\mathbb{R}_{+}$, that is, suppose that the assumptions of Lemma 1 hold. Suppose moreover that $\pi_{1}$ has the $k$th derivative bounded on $\mathbb{R}_{+}$. Let $C > (b-a)^{-1/2}$ and $\alpha > 1$, and define $\tilde{C}(\alpha, C)$ as in (20). Then if $k > \alpha$,

$$
\inf_{f \in \mathcal{S}_{\alpha}} \sup_{f \in \tilde{C}(\alpha,C) \cap \mathbb{H}_{1}} \mathbb{E}_{\pi_{1}} \| \hat{f} - f \|_{\mathbb{H}}^{2} \geq n^{-2\alpha/(k-\alpha)}(1 + o(1)).
$$

Proof. We proceed as in the proof of Theorem 4, that is, we observe that Assumption 3 holds with the same choice of $(g_{k})$ as in Lemma 1 and apply Corollary 2 with $w = 1_{[a,b]}$. Here, the lower bound given in (32) is obtained by showing that

$$
v_{m_{n}} = O(n^{-1} m_{n}^{-\alpha}) \quad \text{as} \quad n \to \infty
$$

holds with $m_{n} = n^{1/(k-\alpha)}$ and with $(v_{m})$ defined by (25). Again we use Lemma 4 to bound $v_{m}$. Here $T$ is the identity operator on $\mathbb{H} = \mathbb{H}'$ and $\pi_{\theta}(x) = \theta^{-1} \pi_{1}(x/\theta)$. Let $M > 0$ be such that the support of $\pi_{1}$ is contained in $[0, M]$. Then for $t \in [a, b]$ and $x > M b$, $\pi_{\theta}(x) = 0$. Hence

$$
\| \pi_{\theta}(x) - P_{m-1}(\pi_{\theta}(x)) \|_{\mathbb{H}} = 0 \quad \text{for all} \quad x > M b.
$$

We now consider the case $x \leq M b$. By the assumption on $\pi_{1}$ and $a$, we have that $\theta \mapsto \pi_{\theta}(x) = \theta^{-1} \pi_{1}(x/\theta)$ is $k$ times differentiable on $[a, b]$. Moreover its $k$th derivative is bounded by $C_{k} x^{k}$ on $[a, b]$, where $C_{k} > 0$ does not depend on $x$. It follows that, for any $h > 0$ and $t \in [a, b]$,

$$
| \Delta^{k}_{h}(\pi_{\theta}(x), t) | \leq C_{k} k! \, (xh)^{k},
$$

MATHEMATICAL METHODS OF STATISTICS Vol. 24 No. 3 2015
where $\Delta^k_h$ is the $k$-th order symmetric difference operator defined by (17). Observing moreover that
\[
\|\pi_n(x)\|_{\mathbb{H}}^2 = \int_a^b t^{-2} \pi_1^2(x/t) \, dt \leq C',
\]
for some $C' > 0$ not depending on $x$, we get that $\pi_n(x) \in C(k, C' \vee C_k \, k! \, x^k)$. Using Corollary 7.25 in Ditzian and Totik [8], we thus have for a constant $C'' > 0$ not depending on $x$,
\[
\|\pi_n(x) - P_{\pi_{n-1}}(\pi_n(x))\|_{\mathbb{H}} \leq C''(1 + x^k)m^{-k} \quad \text{for all } x \leq Mb.
\]
Applying Lemma 4 with (34) and (35), we obtain $v_m = O(m^{-k})$. We conclude that (33) holds with $m_n = n^{1/(k-\alpha)}$, which completes the proof. \hfill $\square$

Theorem 6 provides polynomial lower bounds of the minimax MISE rate, whereas Corollary 1 gives logarithmic upper bounds in the same smoothness spaces. Hence the question of the minimax rate is left completely open in this case. Moreover the lower bound relies on smoothness conditions on $\pi_1$ which rule out Example 3 (for which $\pi_1$ is discontinuous). On the other hand, the case of scale families can be related with the deconvolution problem that has received a considerable attention in a series of papers in the 1990’s (see, e.g., [9–11, 23, 29]). The following section sheds a light on this relationship.

### 5.6. Scale Families and Deconvolution

The following lower bound is obtained from classical lower bounds in the deconvolution problem derived in [11].

**Theorem 7.** Consider the case of scale mixtures on $\mathbb{R}_+$, that is, suppose that Assumption 1 holds with $\zeta$ equal to the Lebesgue measure on $\mathbb{R}_+, \Theta = [a, b] \subset (0, \infty)$ and $\pi_\theta(x) = \theta^{-1} \pi_1(x/\theta)$. Denote by $\phi$ the characteristic function of the density $e^{\theta} \pi_1(e^{\theta})$ on $\mathbb{R}$,
\[
\phi(\xi) = \int e^{i\xi \theta} \pi_1(e^{\theta}) \, d\theta.
\]
Define $\tilde{T}$ as the operator $\tilde{T}(g) = \tilde{g}$, where $g: \mathbb{R} \to \mathbb{R}$ and $\tilde{g}(t) = t^{-1} g(\log(t))$ for all $t \in (0, \infty)$. Let $C > 0$ and $\alpha > 0$, and define $\mathcal{E}(\alpha, C)$ as the set containing all densities $g$ on $\mathbb{R}$ such that
\[
|g^{(r)}(t) - g^{(r)}(u)| \leq C|t - u|^\alpha - r \quad \text{for all } t, u \in \mathbb{R},
\]
where $r = [\alpha]$.

(a) Assume that $\phi^{(j)}(t) = O(|t|^{-\beta - j})$ as $|t| \to \infty$ for $j = 0, 1, 2$, where $\phi^{(j)}$ is the $j$th derivative of $\phi$. Then there exists $C^* > 0$ such that
\[
\inf_{f \in \mathcal{S}_n} \sup_{\tilde{T} \in \mathcal{E}(\alpha, C)} \mathbb{E}_{\pi_1}[\|	ilde{f} - f\|_{\mathbb{H}}^2 \geq C^* n^{-2\alpha/(2(\alpha + \beta) + 1)} (1 + o(1)).
\]

(b) Assume that $\phi(t) = O(|t|^{\beta_0} e^{-|t|^{\beta_0}/\gamma})$ as $|t| \to \infty$ for some $\beta, \gamma > 0$ and $\beta_0$, and that $\pi_1(u) = o(u^{-1} \log(u)^{-\gamma})$ as $u \to 0$, $\infty$ for some $a > 1$. Then there exists $C^* > 0$ such that
\[
\inf_{f \in \mathcal{S}_n} \sup_{\tilde{T} \in \mathcal{E}(\alpha, C)} \mathbb{E}_{\pi_1}[\|	ilde{f} - f\|_{\mathbb{H}}^2 \geq C^* \log(n)^{-2\alpha/\beta} (1 + o(1)).
\]

**Proof.** In the scale mixture case the observation $X$ can be represented as $X = \theta Y$, where $Y$ and $\theta$ are independent variables having density $\pi_1$ and (unknown) density $f$, respectively. By taking the log of the observations, the problem of estimating the density of $\log(\theta)$, that is $f^*(t) = e^x f(e^x)$, is a deconvolution problem. Hence we may apply Theorem 2 in [11] to obtain lower bounds on the nonparametric estimation of $f^*$ from $\log(X_1), \ldots, \log(X_n)$ under appropriate assumptions on $\phi$, which is the characteristic function of $\log(Y)$. Let $a' = \log(a)$ and $b' = \log(b)$. The lower bounds in (a) and (b) above are those appearing in (a) and (b) in Theorem 2 in [11] on the minimax quadratic risk in $\mathbb{H}' = L^2([a', b'])$ for
estimating \( f^* \) in the Lipschitz smoothness class \( \mathcal{L}(\alpha, C) \). Observe that \( \tilde{T} \) is defined for all functions \( g: [a', b'] \to \mathbb{R} \) by \( \tilde{T}(g) = \hat{g} \) with \( \hat{g} \) defined on \([a, b]\) by \( \hat{g}(t) = t^{-1}g(\log(t)) \), so that \( \tilde{T}(f^*) = f \). Observing that \( \tilde{T} \) is a linear operator and that for any \( g \in \mathbb{H}' \), \( \| \tilde{T}(g) \|_{\mathbb{H}} \asymp \| g \|_{\mathbb{H}'} \), we obtain the lower bounds given in (36) and (37).

As in Theorem 6, the smoother \( \pi_1 \) is assumed, the slower the lower bound of the minimax rate. However the lower bounds obtained in Theorem 7 hold for a much larger class of scale families. Indeed, if \( \pi_1 \) is compactly supported, the conditions imposed on \( \pi_1 \) in case (a) are much weaker than in Theorem 6. For instance, it holds with \( \beta = k \) for Example 3. For an infinitely differentiable \( \pi_1 \) both theorems say that the minimax rate is slower than any polynomial rate. However, in this case, case (b) in Theorem 7 may provide a more precise logarithmic lower bound. It is interesting to note that, as a consequence of [5], the MISE rate \( (\log n)^{-2\alpha} \), which is the rate obtained in Corollary 1 by the polynomial estimator for any compactly supported \( \pi_1 \), is the slowest possible minimax rate obtained in Theorem 7(b) for a compactly supported \( \pi_1 \). Such a comparison should be regarded with care since the smoothness class in the latter theorem is different and cannot be compared to the smoothness classes considered in the previous results, as we explain hereafter.

The arguments for adapting the lower bounds of Theorem 7 also apply for minimax upper bounds. More precisely, using the kernel estimators for the deconvolution problem from the observations \( \log(X_1), \ldots, \log(X_n) \) and mapping the estimator through \( \tilde{T} \), one obtains an estimator of \( f \) achieving the same integrated quadratic risk. The obtained rates depend on similar assumptions on \( \phi \) as in (a) and (b), see [9–11]. Although the scale mixture and the convolution model are related to one another by taking the exponential (or the logarithm in the reverse sense) of the observations, it is important to note that, except for Theorem 7, our results are of different nature. Indeed, the upper and lower bounds in the deconvolution problem cannot be compared with those obtained previously in the paper because there are no possible inclusions between the smoothness classes considered in the deconvolution problem and those defined by polynomial approximations.

Let us examine more closely the smoothness class \( \tilde{T}(\mathcal{L}(\alpha, C)) \) that appears in the lower bounds of Theorem 7, inherited from the results on the deconvolution problem. This class contains densities with noncompact supports, whereas \( \tilde{\mathcal{C}}(\alpha, C) \cap \mathbb{H}_1 \) only contains densities with supports in \([a, b]\). Hence neither (36) nor (37) can be used for deriving minimax rates in \( \tilde{\mathcal{C}}(\alpha, C) \cap \mathbb{H}_1 \). In fact the densities exhibited in [11] to prove the lower bound have infinite support by construction and the argument does not at all seem to be adaptable for a class of compactly supported densities. As for upper bounds in the deconvolution problem, they are based on Lipschitz or Sobolev type of smoothness conditions which are not compatible with compactly supported densities on \([a, b]\) except for those that are smoothly decreasing close to the end points. This follows from the fact that, in the deconvolution problem, standard estimators (kernel or wavelet) highly rely on the Fourier behavior both of the mixing density and of the additive noise density. In contrast, such boundary constraints are not necessary for densities in \( \tilde{\mathcal{C}}(\alpha, C) \). For instance, the uniform density on \([a, b]\) belongs to \( \tilde{\mathcal{C}}(\alpha, C) \) for all \( \alpha > 0 \) and \( C > (b - a)^{-1/2} \), but has a Fourier transform decreasing very slowly. A natural conclusion from this observation is that polynomial estimators should be used preferably to standard deconvolution estimators when the mixing density has a known compact support \([a, b] \subset (0, \infty) \). Of course this conclusion holds for both deconvolution and scale mixture problems.

6. SUPPORT ESTIMATION

A basic assumption of our estimation approach is that the mixing density \( f \) belongs to \( \mathbb{H} = L^2[a, b] \). However, in practice the exact interval \([a, b]\) is generally unknown. To cope with this problem, we propose an estimator of the support of the mixing density \( f \), or more precisely of the support of \( Tf \). It can be shown that the support estimator is consistent when it is based on an estimator \( \tilde{T}f_{n,m_n} \), which is a polynomial, and \( Tf \) behaves as follows on the bounds of the support interval. Note that the problem of dealing with an unknown support also occurs in classical density estimation, see [13, 17, 24].
Denote by $[a_0, b_0]$ the smallest interval such that $Tf(u) = 0$ for all $u \in [a', b'] \setminus [a_0, b_0]$. In other words, $a_0 = \inf\{u \in [a', b'], Tf(u) > 0\}$ and $b_0 = \sup\{u \in [a', b'], Tf(u) > 0\}$. Furthermore, we suppose that there exist constants $D > a_0, E < b_0, D', E', \alpha' > 0$ such that

$$Tf(u) \geq \frac{(u - a_0)/D'}{\alpha'} \quad \text{for all } u \in [a_0, D],$$

$$Tf(u) \geq \frac{(b_0 - u)/E'}{\alpha'} \quad \text{for all } u \in [E, b_0].$$

(38) (39)

For fixed $\varepsilon_n, \eta_n > 0$, we define the estimators $\hat{a}_n$ and $\hat{b}_n$ of the interval bounds $a_0$ and $b_0$ by

$$\hat{a}_n = \inf \left\{ u \in [a', b']: T\hat{f}_{n,m}(v) > \frac{\varepsilon_n}{2} \text{ for all } v \in [u, u + \eta_n] \right\},$$

(40)

$$\hat{b}_n = \sup \left\{ u \in [a', b']: T\hat{f}_{n,m}(v) > \frac{\varepsilon_n}{2} \text{ for all } v \in [u - \eta_n, u] \right\}.$$  

(41)

Roughly, these estimators take the smallest and largest value where the estimator $T\hat{f}_{n,m}$ exceeds $\varepsilon_n/2$, by disregarding side-effects of size $\eta_n$. For a suitable choice of the sequences $(\varepsilon_n)$ and $(\eta_n)$ these estimators are consistent.

**Proposition 2.** Let $\hat{f}_{n,m}$ be the density estimator defined in (3) under Assumption 3 with $\alpha > 1/2$. Suppose that $f$ verifies (38)–(39) for appropriate constants $D > a_0, E < b_0, D', E', \alpha' > 0$. Assume that there are sequences $m_n \to \infty, \varepsilon_n \to 0$ and $\eta_n \to 0$ such that

$$\mathbb{E}\|\hat{f}_{n,m} - f\|_\mathcal{H}^2 = O(m_n^{-2\alpha}), \quad \varepsilon_n^{-1} = o(m_n^{2(\alpha - 1)/(2 + 1/\alpha')}), \quad \eta_n = O(m_n^{-1/2}).$$

Then the estimators $\hat{a}_n$ and $\hat{b}_n$ defined by (40) and (41) are consistent for the support bounds $a_0$ and $b_0$. More precisely, as $n \to \infty$,

$$(\hat{a}_n - a_0)_+ = O_P(\varepsilon_n^{1/\alpha'}) \quad \text{and} \quad (\hat{a}_n - a_0)_- = O_P(\varepsilon_n^{1/\alpha'} m_n^{-1}),$$

$$(\hat{b}_n - b_0)_+ = O_P(\varepsilon_n^{1/\alpha'} m_n^{-1}) \quad \text{and} \quad (\hat{b}_n - b_0)_- = O_P(\varepsilon_n^{1/\alpha'}).$$

**Proof.** First we consider $(\hat{a}_n - a_0)_+$. We set $\delta_n = MD\varepsilon_n^{1/\alpha'}$ for some $M > 1$ and denote

$$A_n = \{(\hat{a}_n - a_0)_+ > \delta_n\} = \{\hat{a}_n > a_0 + \delta_n\}$$

$$= \left\{ \forall u \in [a', a_0 + \delta_n] \exists v \in [u, u + \eta_n] \text{ such that } T\hat{f}_{n,m}(v) \leq \frac{\varepsilon_n}{2} \right\}.$$  

As $T\hat{f}_{n,m}$ is a polynomial of degree $m_n$, $T\hat{f}_{n,m}$ has at most $m_n$ intersections with any constant function. Hence the number of subintervals of $[a', b']$, where $T\hat{f}_{n,m}$ exceeds $\varepsilon/2$, is bounded by $m_n$ for any fixed $\varepsilon > 0$. On $A_n$, all such intervals included in $[a', a_0 + \delta_n]$ are at most of size $\eta_n$. Thus, on $A_n$,

$$\int_{a'}^{a_0 + \delta_n} \mathbb{I}\left\{ T\hat{f}_{n,m}(u) > \frac{\varepsilon_n}{2} \right\} du \leq m_n \eta_n.$$  

It follows that on $A_n$

$$\int_{a'}^{a_0 + \delta_n} \mathbb{I}\left\{ T\hat{f}_{n,m}(u) \leq \frac{\varepsilon_n}{2} \right\} du \geq a_0 + \delta_n - a' - m_n \eta_n,$$

and thus

$$\int_{a_0 + D\varepsilon_n^{1/\alpha'}}^{a_0 + D'\varepsilon_n^{1/\alpha'}} \mathbb{I}\left\{ T\hat{f}_{n,m}(u) \leq \frac{\varepsilon_n}{2} \right\} du \geq \delta_n - m_n \eta_n - D'\varepsilon_n^{1/\alpha'}.$$  

For large $n$ such that $\varepsilon_n^{1/\alpha'} < D$ and since $Tf > \varepsilon_n$ on $[a_0 + D'\varepsilon_n^{1/\alpha'}, D]$ by (38), we obtain on $A_n$

$$\int_{a_0 + D'\varepsilon_n^{1/\alpha'}}^{a_0 + D\varepsilon_n^{1/\alpha'}} \mathbb{I}\left\{ T\hat{f}_{n,m}(u) \leq \frac{\varepsilon_n}{2} \right\} du \leq \int_{a_0 + D'\varepsilon_n^{1/\alpha'}}^{a_0 + D\varepsilon_n^{1/\alpha'}} \mathbb{I}\left\{ |T\hat{f}_{n,m}(u) - Tf(u)| > \frac{\varepsilon_n}{2} \right\} du$$

$$\leq \frac{4}{\varepsilon_n^2} \int_{a_0 + D'\varepsilon_n^{1/\alpha'}}^{a_0 + D\varepsilon_n^{1/\alpha'}} (T\hat{f}_{n,m}(u) - Tf(u))^2 du \leq \frac{4}{\varepsilon_n^2} \|T\hat{f}_{n,m} - Tf\|_\mathcal{H}^2 = \frac{4}{\varepsilon_n^2} \|f_{n,m} - f\|_\mathcal{H}^2.$$  

MATHEMATICAL METHODS OF STATISTICS Vol. 24 No. 3 2015
For sufficiently large $M$ we have $m_n \eta_n < \delta_n - D' \varepsilon_n^{1/\alpha'}$. Then it follows by Markov's inequality that
\[
\mathbb{P}(\hat{a}_n - a_0 > \delta_n) \leq \mathbb{P}\left(\frac{4}{\varepsilon_n^2} \left\| \hat{f}_{n,m_n} - f \right\|_{\mathbb{H}}^2 \geq \delta_n - m_n \eta_n - D' \varepsilon_n^{1/\alpha'}\right) \\
\leq \frac{4 \mathbb{E}(\|\hat{f}_{n,m_n} - f\|_{\mathbb{H}}^2)}{\varepsilon_n^2 (\delta_n - m_n \eta_n - D' \varepsilon_n^{1/\alpha'})} \to 0, \quad n \to \infty,
\]
by the assumptions on $(\varepsilon_n)_n$ and $\mathbb{E}(\|\hat{f}_{n,m_n} - f\|_{\mathbb{H}}^2)$ and as $\delta_n = M D' \varepsilon_n^{1/\alpha'}$. Thus $(\hat{a}_n - a_0) = O_P(\varepsilon_n^{1/\alpha'})$.

To investigate $(\hat{a}_n - a_0)_-$ put $\delta_n = M' \eta_n$ for some $M' > 1$. By using that $Tf = 0$ on $[a, a_0]$, we have
\[
\mathbb{P}(\hat{a}_n - a_0 > \delta_n) = \mathbb{P}(\hat{a}_n < a_0 - \delta_n) \\
= \mathbb{P}\left(\exists x \in [a', a_0 - \delta_n]: \int_x^{x + \eta_n} 1\left\{ T\hat{f}_{n,m_n}(u) > \frac{\varepsilon_n}{2} \right\} du = \eta_n\right) \\
\leq \mathbb{P}\left(\int_{a'}^{a_0} 1\left\{ T\hat{f}_{n,m_n}(u) > \frac{\varepsilon_n}{2} \right\} du \geq \eta_n\right) = \mathbb{P}\left(\int_{a'}^{a_0} 1\left\{ |T\hat{f}_{n,m_n}(u) - Tf(u)|^2 > \frac{\varepsilon_n^2}{4} \right\} du > \eta_n\right) \\
\leq \mathbb{P}\left(\frac{4}{\varepsilon_n^2} \int_{a'}^{a_0} |T\hat{f}_{n,m_n}(u) - Tf(u)|^2 du > \eta_n\right) \leq \frac{4 \mathbb{E}(\|\hat{f}_{n,m_n} - f\|_{\mathbb{H}}^2)}{\eta_n \varepsilon_n^2} \to 0,
\]
where again we applied Markov’s inequality. Consequently, $(\hat{a}_n - a_0)_- = O_P(\eta_n) = O_P(\varepsilon_n^{1/\alpha'}/m_n^{1/\alpha'})$.

By symmetry, the properties on $\hat{b}_n$ stated in the proposition hold as well. \qed

By Theorem 1 the proposition applies to Examples 1 (a) and 3 with $m_n = A \log n$ and to Examples 1 (b) and 2 with $m_n = A \log n / \log \log n$.

7. NUMERICAL RESULTS

A simulation study is conducted to evaluate the performance of the estimator on finite datasets. Six different mixture settings are considered, namely the exponential mixture from Example 1 (a) and 1 (b), the Gamma shape mixture from Example 2, the uniform mixture and the Beta mixture with $k = 4$ from Example 3 and the exponential mixture with a location parameter from Example 4.

We consider the case where the mixing density $f$ is the Beta distribution on the interval $[1, 4]$ with parameters $\alpha = 3/2$ and $\beta = 3$. Remark that for the exponential mixture setting of Example 1 (b) we cannot take a mixing distribution with support $[a, b]$ with $a = 0$, since $b' = 1/a$ must be finite.

For every mixture setting, the estimator $\hat{f}_{n,m_n}$ with $m = 5$ is computed on a large number of datasets (for sample sizes $n$ varying from 100 to $10^9$) and the corresponding MISE is evaluated. Table 1 gives the mean values of the different MISE’s and the associated standard deviations. Obviously, in all six settings the MISE decreases when $n$ increases. Note that in the last four settings, where the mixing density $f$ is approximated in the same polynomial basis, the MISE tends to the same value, which is obviously the squared bias of the estimator when $m = 5$. In the exponential mixture settings, different values are obtained because different bases are used to approach $f$. The exponential mixture setting from Example 1 (a) always has the largest mean MISE value, while the uniform and the Beta mixtures are doing best.

Figure 1 illustrates the estimator $\hat{f}_{n,m_n}$ when $n = 10^5$ and where the order $m$ is the value minimizing the MISE when $n = 10^5$, say $m_{\text{best}}$. The values of $m_{\text{best}}$ have been obtained by extra simulations. We see that in the first two settings, we only have $m_{\text{best}} = 3$ and the estimator seems slightly biased. On the contrary, the uniform mixture setting allows for the best approximation with $m_{\text{best}} = 5$. 

MATHEMATICAL METHODS OF STATISTICS Vol. 24 No. 3 2015
Table 1. Estimated MISE (and standard deviation) of estimator \( \hat{f}_{m,n} \) with \( m = 5 \) in six different mixture settings when the mixing density \( f \) is a Beta distribution

|          | \( n \) | \( 10^2 \) | \( 10^3 \) | \( 10^4 \) | \( 10^5 \) | \( 10^6 \) | \( 10^7 \) | \( 10^8 \) | \( 10^9 \) |
|----------|--------|----------|----------|----------|----------|----------|----------|----------|----------|
| **Exp. (a)** | MISE  | 0.72     | 0.69     | 0.62     | 0.48     | 0.26     | 0.058    | 8.4e-03  | 1.1e-03  |
|          | sd     | (0.16)   | (0.17)   | (0.16)   | (0.19)   | (0.18)   | (0.085)  | (0.10)   | (1.4e-03) |
| **Exp. (b)** | MISE  | 0.61     | 0.52     | 0.35     | 0.21     | 0.084    | 0.015    | 2.0e-03  | 4.4e-04  |
|          | sd     | (0.21)   | (0.25)   | (0.25)   | (0.21)   | (0.12)   | (0.027)  | (2.9e-03) | (2.7e-04) |
| **Gamma** | MISE   | 0.58     | 0.47     | 0.31     | 0.12     | 0.020    | 3.4e-03  | 1.5e-03  | 1.3e-03  |
|          | sd     | (0.20)   | (0.22)   | (0.21)   | (0.13)   | (0.024)  | (3.0e-03) | (4.5e-04) | (4.9e-05) |
| **Uniform** | MISE  | 0.32     | 0.10     | 0.015    | 2.9e-03  | 1.5e-03  | 1.3e-03  | 1.3e-03  | 1.3e-03  |
|          | sd     | (0.25)   | (0.12)   | (0.018)  | (2.1e-03) | (2.6e-04) | (4.4e-05) | (1.2e-05) | (3.6e-06) |
| **Beta**  | MISE   | 0.46     | 0.19     | 0.035    | 5.4e-03  | 1.7e-03  | 1.4e-03  | 1.3e-03  | 1.3e-03  |
|          | sd     | (0.27)   | (0.17)   | (0.040)  | (5.3e-03) | (6.5e-04) | (8.6e-05) | (2.2e-05) | (5.8e-06) |
| **Exp. loc.** | MISE  | 0.55     | 0.47     | 0.29     | 0.11     | 0.015    | 2.9e-03  | 1.5e-03  | 1.3e-03  |
|          | sd     | (0.20)   | (0.23)   | (0.21)   | (0.11)   | (0.018)  | (1.9e-03) | (2.6e-04) | (5.3e-05) |

Fig. 1. 10 estimators \( \hat{f}_{m_{\text{best}},n} \) in six different settings with \( n = 10^5 \) when the mixing density \( f \) is a Beta distribution (dashed curve).

(a) Exp. mixture (a), \( m_{\text{best}} = 3 \)  (b) Exp. mixture (b), \( m_{\text{best}} = 3 \)  (c) Gamma mixture, \( m_{\text{best}} = 4 \)
(d) Uniform mixture, \( m_{\text{best}} = 5 \)  (e) Beta mixture, \( m_{\text{best}} = 4 \)  (i) Exp. location, \( m_{\text{best}} = 4 \)
APPENDIX A: TECHNICAL RESULTS

Lemma 5. Let $\alpha > 0$, $a < b$ and $a' < b'$. Define $\tilde{C}(\alpha, C)_{\mathbb{H}}$ as in (20) and $\tilde{C}'(\alpha, C)_{\mathbb{H}}$ similarly with $a'$ and $b'$ replacing $a$ and $b$. Let $\sigma$ be $[\alpha] + 1$ differentiable on $[a, b]$ and $\tau': [a', b'] \rightarrow [a, b]$ be $[\alpha] + 1$ differentiable on $[a', b']$ with a nonvanishing first derivative. Then

$$\{ \sigma f: f \in \tilde{C}(\alpha, \cdot)_{\mathbb{H}} \} \hookrightarrow \tilde{C}(\alpha, \cdot)_{\mathbb{H}} \quad \text{and} \quad \{ f \circ \tau: f \in \tilde{C}(\alpha, \cdot)_{\mathbb{H}} \} \hookrightarrow \tilde{C}'(\alpha, \cdot)_{\mathbb{H}}.$$

Proof. As the first embedding is the inclusion (40) in Roueff and Ryden [25], we only show the second embedding. Let $f \in \tilde{C}(\alpha, C)$ and denote $r = [\alpha] + 1$. Let $t \in (0, 1]$. By the equivalence (19) with the $K$-functional given in (18) there exists a function $h$ such that $h^{(r-1)} \in A.C._{\text{loc}}$ and

$$\|f - h\|_{H} + t^{r}\|\varphi^{r}h^{(r)}\|_{H} \leq 2M\omega_{\varphi}(f, t)_{\mathbb{H}} \leq 2MCt^{\alpha},$$

(42)

where $\varphi(x) = \sqrt{(x-a)(b-x)}$. Let us set $\tilde{h} = h \circ \tau$ and show that, for some constant $K > 0$ depending neither on $t$ nor on $C$,

$$\|f \circ \tau - \tilde{h}\|_{H^r} + t^{r}\|\varphi^{r}\tilde{h}^{(r)}\|_{H^r} \leq K Ct^{\alpha},$$

(43)

where we defined $\tilde{\varphi}(x) = \sqrt{(x-a')(b'-x)}$, that is the same definition as $\varphi$ with $a'$ and $b'$ replacing $a$ and $b$. Using again equivalence (19), the bound in (43) will achieve the proof of the lemma.

Note that since $\tau'$ does not vanish, denoting $C_{1} = (\inf |\tau'|)^{-1}$, for all $g \in \mathbb{H}$, we have

$$\|g \circ \tau\|_{H^{r}} \leq C_{1}\|g\|_{H}.$$  

(44)

In particular, we have that

$$\|f \circ \tau - \tilde{h}\|_{H^{r}} = \|(f - h) \circ \tau\|_{H^{r}} \leq C_{1}\|f\|_{H}.$$  

(45)

Since $\tau$ is $r$ times continuously differentiable and $h^{(r-1)} \in A.C._{\text{loc}}$, we note that $\tilde{h}^{(r-1)} \in A.C._{\text{loc}}$ with $\tilde{h}^{(r)} = \sum_{j=1}^{r} \tau_{j} \times h^{(j)} \circ \tau$, where the $\tau_{j}$’s are continuous functions only depending on $\tau$. Hence there is a constant $C_{2} > 0$ only depending on $\tau$ and $r$ such that

$$\|\varphi^{r}\tilde{h}^{(r)}\|_{H^{r}} \leq C_{2}\max_{j=1, \ldots, r} \|\varphi^{r}h^{(j)} \circ \tau\|_{H^{r}}.$$  

Another simple consequence of $\tau'$ not vanishing on $[a', b']$ is that there exists a constant $C_{3} > 0$ such that $\tilde{\varphi}(x) \leq C_{3}\varphi \circ \tau(x)$ for all $x \in [a', b']$. Using this with (44) in the previous display, we get

$$\|\varphi^{r}\tilde{h}^{(r)}\|_{H^{r}} \leq C_{1}C_{2}C_{3}\max_{j=1, \ldots, r} \|\varphi^{r}h^{(j)}\|_{H}.$$  

(46)

We shall prove that $\|\varphi^{r}h^{(j)}\|_{H}$ appearing in the right-hand side of the previous inequality is in fact maximized, up to multiplicative and additive constants, at $j = r$. For $j = 1, \ldots, r - 1$, we proceed recursively as follows. For any $u \in (a, b)$, we have

$$|h^{(j)}(x)| \leq \left| \int_{u}^{x} h^{(j+1)}(s) \, ds \right| + |h^{(j)}(u)|.$$  

Then, by Jensen’s inequality,

$$\|\varphi^{r}h^{(j)}\|_{H} \leq \left\{ \int_{x=a}^{b} \varphi^{2r}(x) \left( |x - u| \int_{s \in [u, x]} \{h^{(j+1)}(s)\}^{2} \, ds \right) \, dx \right\}^{1/2} + \|\varphi^{r}\|_{H} \|h^{(j)}(u)\|,$$

where we used the convention that $[c, d]$ denotes the same segment for $c \leq d$ as well as for $c \geq d$. By Fubini’s theorem, the term between braces reads

$$\int_{s=a}^{b} \{h^{(j+1)}(s)\}^{2} \psi(s; u) \, ds \quad \text{with} \quad \psi(s; u) = \int_{[u, x]}(s) (x - a)^{r}(b - x)^{r} |x - u| \, dx.$$
Let \( \tilde{a} < \tilde{b} \) be two fixed numbers in \((a, b)\). It is straightforward to show that, for some constant \( C_4 > 0 \) only depending on \( a, b, \tilde{a}, \tilde{b} \), we have
\[
\psi(s; u) \leq C_4^2 \varphi^{2r}(s) \quad \text{for all } u \in (\tilde{a}, \tilde{b}).
\]

The last three displays thus give that
\[
\|\varphi^r h^{(j)}\|_{\mathcal{H}} \leq C_4 \|\varphi^r h^{(j+1)}\|_{\mathcal{H}} + \|\varphi^r\|_{\mathcal{H}} \inf_{u \in [\tilde{a}, \tilde{b}]} |h^{(j)}(u)|.
\]

By induction in \( j \), we thus get with (46) that there is a constant \( C_5 \) such that
\[
\|\varphi^r \tilde{h}^{(r)}\|_{\mathcal{H}^r} \leq C_5 \left( \|\varphi^r h^{(r)}\|_{\mathcal{H}} + \sum_{j=1, \ldots, r-1} \inf_{u \in [\tilde{a}, \tilde{b}]} |h^{(j)}(u)| \right).
\]

The final step of the proof consists in bounding \( \delta_j = \inf_{u \in [\tilde{a}, \tilde{b}]} |h^{(j)}(u)| \) for \( j = 1, \ldots, r - 1 \). For any \( v, v' \in [\tilde{a}, \tilde{b}] \), we have \( |h^{(j-1)}(v') - h^{(j-1)}(v)| \geq \delta_j |v' - v| \). Suppose that \( v \) is in the first third of the segment \([\tilde{a}, \tilde{b}]\) and \( v' \) in the last third so that \(|v - v'| \geq (\tilde{b} - \tilde{a})/3 \). On the other hand, \(|h^{(j-1)}(v') - h^{(j-1)}(v)| \leq |h^{(j-1)}(v')| + |h^{(j-1)}(v)|\). It follows that \( |h^{(j-1)}(v')| \) and \( |h^{(j-1)}(v)| \) cannot be both less than \( \delta_j (\tilde{b} - \tilde{a})/3 \), which provides a lower bound of \(|h^{(j-1)}|\) on at least one sub-interval of \([\tilde{a}, \tilde{b}]\) of length \((\tilde{b} - \tilde{a})/3\). Proceeding recursively we get that there exists a sub-interval of \([\tilde{a}, \tilde{b}]\) on which \( \tilde{h} \) is lower bounded by \( \delta_j \) multiplied by some constant. This in turn gives that
\[
\sum_{j=1, \ldots, r-1} \inf_{u \in [\tilde{a}, \tilde{b}]} |h^{(j)}(u)| \leq C_6 \|h\|_{\mathcal{H}},
\]

where \( C_6 \) is a constant only depending on \( \tilde{a}, \tilde{b} \) and \( r \). Observe that, since \( f \in \tilde{C}(\alpha, C) \), we have \( \|f\|_{\mathcal{H}} \leq C \). Using (42), \( t \in (0, 1] \) and \( \|h\|_{\mathcal{H}} \leq \|f - h\|_{\mathcal{H}} + \|f\|_{\mathcal{H}} \) in the last display we thus get
\[
\sum_{j=1, \ldots, r-1} \inf_{u \in [\tilde{a}, \tilde{b}]} |h^{(j)}(u)| \leq C_6 (2M + 1) C.
\]

Finally, this bound, (47), (45) and (42) yield (43) and the proof is achieved. \( \square \)

**Lemma 6.** Let \((p_k)\) be the sequence of polynomials defined by \( p_1(t) = 1 \), \( p_2(t) = t \), \( \ldots, p_k(t) = t(t + 1) \ldots (t + k - 2) \) for all \( k \geq 2 \). Define the coefficients \( \tilde{c}_{k, l} \) by the expansion formula
\[
t^{k-1} = \sum_{l=1}^{k} \tilde{c}_{k, l} p_l(t) \quad \text{valid for } k = 1, 2, \ldots. \quad \text{Then } \tilde{c}_{1, 1} = 1, \text{ and for all } k \geq 2, \quad \tilde{c}_{k, 1} = 0, \quad \tilde{c}_{k, k} = 1 \quad \text{and} \quad \tilde{c}_{k, l} = \tilde{c}_{k-1, l-1} - (l - 1) \tilde{c}_{k-1, l} \quad \text{for all } l = 2, \ldots, k - 1. \]

Moreover, we have, for all \( k \geq 1, \)
\[
\sum_{l=1}^{k} |\tilde{c}_{k, l}| \leq k!.
\]

**Proof.** By definition of \( p_l \), we have \( tp_l(t) = p_{l+1}(t) - (l - 1) p_l(t) \) for any \( l \geq 1 \). Hence, for any \( k \geq 2, \)
writing \( t^{k-1} = tt^{k-2} = \sum_{l=1}^{k} \tilde{c}_{k-1, l} t p_l(t), \) we obtain (48).

We now prove (49). It is obviously true for \( k = 1. \) From (48), it follows that, for all \( k \geq 2, \)
\[
\sum_{l=1}^{k} |\tilde{c}_{k, l}| \leq \sum_{l=1}^{k-1} l |\tilde{c}_{k-1, l}| + 1.
\]

Bounding \( l \) inside the last sum by \((k - 1)\) yields (49). \( \square \)
REFERENCES

1. K. E. Andersen and M. B. Hansen “Multiplicative Censoring: Density Estimation by a Series Expansion Approach”, J. Statist. Plann. Inference 98, 137–155 (2001).
2. M. Asgharian, M. Carone, and V. Fakoor, “Large-Sample Study of the Kernel Density Estimators under Multiplicative Censoring”, Ann. Statist. 40 (1), 159–187 (2012).
3. F. Baladandou and J. A. Wellner, “Estimation of a k-Monotone Density: Limit Distribution Theory and the Spline Connection”, Ann. Statist. 35 (6), 2536–2564 (2007).
4. D. Belomestny and J. Schoenmakers, Statistical Skorohod Embedding Problem and Its Generalizations, Technical Report, 2014. URL http://arxiv.org/abs/1407.0873.
5. A. Beurling and P. Malliavin, “On Fourier Transforms of Measures with Compact Support”, Acta Math. 107, 291–309 (1962).
6. F. Comte and Y. Genon-Catalot, “Adaptive Laguerre Density Estimation for Mixed Poisson Models”, Electronic J. Statist. 9, 1112–1148 (2015).
7. R. A. DeVore and G. G. Lorentz, Constructive Approximation (Springer, 1993).
8. Z. Ditzian and V. Totik, Moduli of Smoothness, in Springer Series in Computational Mathematics (Springer, 1987).
9. Jianqing Fan, “On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems”, Ann. Statist. 19 (3), 1257–1272 (1991).
10. Jianqing Fan, “Global Behavior of Deconvolution Kernel Estimates”, Statist. Sinica 1 (2), 541–551 (1991).
11. Jianqing Fan, “Adaptively Local One-Dimensional Subproblems with Application to a Deconvolution Problem”, Ann. Statist. 21 (2), 600–610 (1993).
12. W. Feller, An Introduction to Probability Theory and Its Applications, 2nd ed. (Wiley, New York, 1971), Vol. 2.
13. G. Gayraud, “Estimation of Functionals of Density Support”, Math. Meth. Statist. 6, 26–47 (1997).
14. C. Goutis, “Nonparametric Estimation of a Mixing Density via the Kernel Method”, J. Amer. Statist. Assoc. 92 (440), 1445–1450 (1997).
15. N. W. Hengartner, “Adaptive Demixing in Poisson Mixture Models”, Ann. Statist. 25 (3), 917–928 (1997).
16. N. P. Jewell, “Mixtures of Exponential Distributions”, Ann. Statist. 10 (2), 479–484 (1982).
17. A. Juditsky and S. Lambert-Lacroix, “On Minimax Density Estimation on R”, Bernoulli 10 (2), 187–220 (2004).
18. P. Kvam, “Length Bias in the Measurements of Carbon Nanotubes”, Technometrics 50 (4), 462–467 (2008).
19. Nan Laird, “Nonparametric Maximum Likelihood Estimation of a Mixing Distribution”, J. Amer. Statist. Assoc. 73 (364), 805–811 (1978).
20. J. R. Lakowicz, Principles of Fluorescence Spectroscopy (Academic/Plenum, New York, 1999).
21. B. G. Lindsay, “The Geometry of Mixture Likelihoods: A General Theory”, Ann. Statist. 11 (1), 86–94 (1983).
22. F. Olmos, B. Kauffmann, A. Simonian, and Y. Carlinet, “Catalog Dynamics: Impact of Content Publishing and Perishing on the Performance of a LRU Cache”, in Proc. 26th Internat. Telettraffic Congress, Karlshkrona, Sweden (2014) URL http://arxiv.org/abs/1403.5479.
23. M. Pensky and B. Vidakovic, “Adaptive Wavelet Estimator for Nonparametric Density Deconvolution”, Ann. Statist. 27 (6), 2033–2053 (1999).
24. P. Reynaud-Bouret, V. Rivoirard, and Ch. Tuleau-Malot, “Adaptive Density Estimation: A Curse of Support?”, J. Statist. Plann. Inference 141, 115–139 (2011).
25. F. Roueff and T. Ryden, “Nonparametric Estimation of Mixing Densities for Discrete Distributions”, Ann. Statist. 33, 2066–2108 (2005).
26. B. Valeur, Molecular Fluorescence (Wiley-VCH, Weinheim, 2002).
27. Y. Vardi, “Multiplicative Censoring, Renewal Processes, Deconvolution and Decreasing Density: Nonparametric Estimation”, Biometrika 76, 751–761 (1989).
28. S. Venturini, F. Dominici, and G. Parmigiani, “Gamma Shape Mixtures for Heavy-Tailed Distributions”, Ann. Appl. Statist. 2 (2), 756–776 (2008).
29. Cun-Hui Zhang, “Fourier Methods for Estimating Mixing Densities and Distributions”, Ann. Statist. 18, 806–831 (1990).
30. Cun-Hui Zhang, “On Estimating Mixing Densities in Discrete Exponential Family Models”, Ann. Statist. 23, 929–945 (1995).