Identifiability and Rates of Estimation for Scale Parameters in Location Mixture Models

Hemant Ishwaran

Annals of Statistics, Volume 24, Issue 4 (Aug., 1996), 1560-1571.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199608%2924%3A4%3C1560%3AIAROEF%3E2.0.CO%3B2-R

The Annals of Statistics
1996, Vol. 24, No. 4, 1560–1571

IDENTIFIABILITY AND RATES OF ESTIMATION FOR SCALE PARAMETERS IN LOCATION MIXTURE MODELS[1]

BY HEMANT ISHWARAN

University of Ottawa

In this paper we consider the problem of identifiability and estimation for the scale parameter θ in the location mixture model $\theta(X + Y)$, where X has a known distribution independent of the Y, whose distribution is unknown. Identification of θ is ensured by constraining Y based on the tail behavior of the distribution for X. Rates for estimation are described for those X which can be written as a square summable series of exponential variables. As a special case, our analysis shows that the structural parameters in the Weibull semiparametric mixture (Heckman and Singer) are not estimable at the usual parametric $O_p(1/\sqrt{n})$ rate. The exact relationship between identifying constraints and achievable rates is explained.

1. Introduction. Heckman and Singer (1984) studied economic theories concerning continuous durations of occupancy of states. In order to properly estimate structural parameters in the presence of population heterogeneity, they proposed the use of a semiparametric mixture model as a method for modeling duration data. The model they proposed was the Weibull semiparametric mixture

$$f(t \mid z, \beta, \theta, G) = \{t > 0\} \int \frac{1}{\theta} t^{1/\theta - 1} \exp(-\beta' z - y - t^{1/\theta} \exp(-\beta' z - y))\, dG(y), \tag{1}$$

where t is the observed positive duration time, z is a vector of time-invariant observed covariates and (β, θ) is a vector of unknown structural parameters in $\mathbb{R}^k \otimes (0, \infty)$. The distribution G in (1) (referred to as the unknown mixing distribution) was introduced in order to account for potential heterogeneity, and was assumed to be completely unspecified up to an identifying moment constraint.

Let Y denote the random variable with unknown distribution G, and let W denote a standard exponential variable. If W and Y are independent, then

$$T = [W \exp(\beta' z + Y)]^{\theta} \tag{2}$$

Received March 1994; revised October 1995.
[1]Research supported in part by NSF Grant DMS-91-02286.
AMS 1991 *subject classifications.* Primary 62G05; secondary 62G20, 62P20.
Key words and phrases. Weibull semiparametric mixture, mixture model, structural parameter.

has the (conditional) Weibull mixture density (1). In the absence of covariate information, or equivalently if we set $\beta = 0$, a log transformation in (2) results in the location mixture

$$\theta(\log W + Y), \tag{3}$$

with unknown scale parameter θ.

In general, if X has known distribution independent of unknown Y, then

$$M = \theta(X + Y), \qquad \theta > 0, \tag{4}$$

describes a location mixture model with unknown scale parameter θ. The intent of this paper is to study the problem of identification and estimation for θ in the general mixture (4). Given a sample of n independent realizations from (4), this paper studies the relationship between identifying constraints on Y and uniform lower rates of estimation for θ. Because the problem of estimation for θ from the Weibull mixture (1) is at least as difficult as from the transformed and reduced model (3), the Weibull mixture will serve as both a special case and as motivation for this problem.

Jewell (1982) observed that a theorem of Bernstein implies the existence of a variable Y, independent of W, such that

$$W^{\theta_1} =_D W^{\theta_2}\exp(Y) \quad \text{where } 0 < \theta_2 < \theta_1.$$

That is, a Weibull distribution with a fixed shape parameter can always be written as a scale mixture of Weibull distributions. However, because Jewell based his argument on an existence proof, no explicit form for Y in the construction was given excepting for the special case $\theta_1 = 2$ and $\theta_2 = 1$.

Jewell's (1984) observation is of particular relevance to this paper, for it shows that the Weibull mixture model (1) is unidentified without constraints. In Section 3 we will provide an explicit construction for Y as a direct proof of the lack of identification. If σ_j denotes a sequence of binomial variables with distribution $\text{Bin}(1, 1 - \theta_2/\theta_1)$, and $W_j \sim \exp(1)$ a sequence of standard exponentials, then, under the assumption that all variables are mutually independent, the nonidentifiability of (3) [and hence (1)] follows from the equality of distributions:

$$\theta_1 \log W =_D \theta_2(\log W + Y_2) \quad \text{for } 0 < \theta_2 < \theta_1, \tag{5}$$

where

$$Y_2 = -\gamma\left(\frac{\theta_1}{\theta_2} - 1\right) - \frac{\theta_1}{\theta_2}\sum_{j=1}^{\infty}\frac{1}{j}\left(\sigma_j W_j - 1 + \frac{\theta_2}{\theta_1}\right), \tag{6}$$

and γ = Euler's constant ≈ 0.57722.

The construction (5) shows that θ need not be identified in the general mixture (4). Section 3 studies the identification problem. There we present Theorem 15 which states conditions on the moments for Y sufficient to ensure identification of θ in (4). The construction for the nuisance variable (6) is especially useful in understanding the conditions and application of the theorem. The construction will enable us to identify exactly which terms in the sum-representation create problems for identification.

Theorem 20 of Section 4 expresses the main result of the paper by describing the relationship between moment constraints on Y (sufficient to ensure identifiability) and lower rates for estimation of θ. The theorem applies to those X variables which can be written as a square summable series of centered exponentials, and consequently is applicable to the transformed Weibull mixture (3). In particular, establishing a lower rate in the constrained mixture (3) involves modifying the variable (6) just enough to satisfy the required moment constraints, while still making it difficult to distinguish between the $\theta_1 \log W$ variable and the $\theta_2(\log W + Y_2)$ mixture. The analysis shows that the scale parameter in the Weibull mixture cannot be estimated at better than an $O_p(n^{-d/(2d+1)})$ rate under a $(1 + d)$th moment constraint to the mixing Y. This seems to be a sharp lower rate. Under a second moment constraint, Honoré (1990) constructed a class of estimators that achieve an $O_p(n^{-s}/\log n)$ rate of estimation for each $0 < s < 1/3$. Preliminary work by Ishwaran (1994) suggests that $O_p(n^{-d/(2d+1)})$ is the optimal rate for $0 < d < 1$, while preliminary results by Honoré (1994) suggest that his class of estimators can be extended to achieve an $O_p(n^{-s}/\log n)$ rate under a $(1 + k)$th constraint for each $0 < s < k/(2k + 1)$, where $k \geq 1$ is any integer.

In summary, the layout of the paper is as follows. Section 2 provides a working definition for locally uniform rates of estimation. Section 3 contains the Weibull construction and describes conditions which ensure identifiability of θ in the general mixture model (4). Section 4 contains the main result in Theorem 20.

A word concerning notation. In most places in the paper, the linear functional notation for expectation is employed. For example, the expected value of a function g with respect to a probability measure $\mathbb{P}$ is written as $\mathbb{P}g$, or $\mathbb{P}g(X)$, rather than the usual convention $\int g(x)\, d\mathbb{P}(x)$. One exception is that $\int g$, or $\int g(x)\, dx$, will be written to denote the integral of g with respect to Lebesgue measure.

2. Locally uniform estimation. The scale–location mixture problem falls under the following general framework. Let $\mathscr{P}$ be a family of probability models on a common measurable space $(\mathscr{X}, \mathscr{A})$, and let λ be the functional which maps a probability $\mathbb{P} \in \mathscr{P}$ onto its structural parameter $\lambda(\mathbb{P}) = \theta$ in the metric space (D, d). Thus, in the context of our mixture problem, $\mathscr{P}$ denotes a family of models consisting of mixtures of the form (4), while $\lambda(\mathbb{P})$ is the functional which maps $\mathbb{P} \in \mathscr{P}$ onto its scale parameter θ in $(0, \infty)$.

Locally uniform estimation by an estimator $\hat{\theta}_n$ for θ will mean that $\hat{\theta}_n$ becomes close to $\lambda(\mathbb{P})$ uniformly for all $\mathbb{P}$ over a (possibly changing) family of models $\mathscr{P}_n \subseteq \mathscr{P}$.

DEFINITION 7. Let $\mathscr{P}_n$ be a sequence of families in $\mathscr{P}$, and let δ_n be a decreasing positive sequence. Estimators $\hat{\theta}_n$ for $\lambda(\mathbb{P})$ are said to have an

$O_p(\delta_n)$ rate of convergence over $\mathscr{P}_n$ if for each $\varepsilon > 0$ there exists a finite constant κ_ε such that

$$\limsup_{n\to\infty} \sup_{\mathbb{P}\in\mathscr{P}_n} \mathbb{P}^n\left\{d\left(\hat{\theta}_n, \lambda(\mathbb{P})\right) \geq \kappa_\varepsilon \delta_n\right\} \leq \varepsilon,$$

where $\mathbb{P}^n$ denotes the n-fold product measure $\mathbb{P} \otimes \cdots \otimes \mathbb{P}$ (n factors) for a probability $\mathbb{P}$.

The crucial idea for determining lower bounds for rates on estimation involves translating a proposed rate of convergence into an assertion involving $\mathscr{L}_1$-distances between probability measures. The technique is originally due to Le Cam (1973), and has more recently been studied by Donoho and Liu (1987, 1991).

LEMMA 8. *Suppose there exist models* $\mathbb{P}_n, \mathbb{Q}_n \in \mathscr{P}_n$ *such that*

$$\limsup_{n\to\infty} \|\mathbb{P}_n^n - \mathbb{Q}_n^n\|_1 < 2,$$

where $\|\cdot\|_1$ *denotes the* $\mathscr{L}_1$*-distance. Then* θ *cannot be estimated at a rate better than* $O_p(\lambda(\mathbb{P}_n, \mathbb{Q}_n))$ *over* $\mathscr{P}_n$ *in the sense of Definition* 7.

PROOF. Let $A_n = \{d(\hat{\theta}_n, \lambda(\mathbb{P}_n)) < \delta_n/2\}$, where $\delta_n = \lambda(\mathbb{P}_n, \mathbb{Q}_n)$ and $\hat{\theta}_n$ is some estimator for θ. Then $\mathbb{Q}_n^n A_n \leq \mathbb{Q}_n^n\{d(\hat{\theta}_n, \lambda(\mathbb{Q}_n)) \geq \delta_n/2\}$, by the δ_n separation between $\mathbb{P}_n$ and $\mathbb{Q}_n$. If $\hat{\theta}_n$ had a rate better than $O_p(\delta_n)$, then the right-hand side of the last inequality would eventually be bounded by arbitrarily small $\varepsilon > 0$, while $\mathbb{P}_n^n A_n \geq 1 - \varepsilon$ eventually. This would lead to the contradiction

$$\tfrac{1}{2}\|\mathbb{P}_n^n - \mathbb{Q}_n^n\|_1 \geq |\mathbb{P}_n^n A_n - \mathbb{Q}_n^n A_n| \geq 1 - 2\varepsilon \quad \text{eventually.}$$

Thus $\hat{\theta}_n$ cannot do better than $O_p(\delta_n)$. □

In order to establish the best possible lower rate in the mixture problem, our strategy will be to search for mixtures $\theta_0(X + Y_0)$ and $\theta_n(X + Y_n)$ which are close in the Hellinger distance, but whose θ parameters are as far apart as possible. The Hellinger calculation is convenient because it provides a simple method for bounding the $\mathscr{L}_1$-distance between n-fold product measures [see, e.g., Le Cam (1973), Lemma 1]. In particular, by Lemma 8, in order to establish a lower rate of $O_p(\theta_n - \theta_0)$ for θ, it would suffice to find mixtures so that

$$H(\theta_0(X + Y_0), \theta_n(X + Y_n)) \leq \frac{\varepsilon}{\sqrt{n}} \quad \text{for small } \varepsilon > 0,$$

where $H(V_1, V_2)$ is the Hellinger distance between any two random variables V_1 and V_2. This is the method used in Section 4 for establishing rates of estimation.

3. Identifiability. Consider the problem of identifiability of the scale parameter θ from knowledge of the distribution of the location mixture model (4), where X has a known fixed distribution F independent of the Y, which itself has unknown distribution G. To motivate the problem, let us first consider scale mixtures of Weibull distributions, that is, distributions of random variables of the form $[W\exp(Y)]^\theta$, where Y is a positive random variable independent of $W \sim \exp(1)$. A direct construction will show that, without constraints on the mixing distribution, the Weibull scale mixtures are not identified.

The construction will use the following representation.

LEMMA 9. *Suppose W_j' is an i.i.d. sequence of* exp(1) *variables. Then there exists a random variable $W \sim \exp(1)$ such that*

$$\log W = -\gamma - \sum_{j=1}^{\infty} (W_j' - 1)/j, \tag{10}$$

where γ = Euler's constant.

PROOF. Let $H_N = \sum_{j=1}^{N} 1/j$. Then $H_N = \log N + \gamma + o(1)$. Define X_N to be the partial sum to N terms of the right-hand side of (10). Use the fact that

$$W_{(N)}' = \max_{1 \le j \le N} W_j' =_D W_N'/N + \cdots + W_1'/1,$$

to see that

$$X_N =_D -\gamma + H_N - W_{(N)}'.$$

It follows that

$$\begin{aligned}
\mathbb{P}\{X_N > x\} &= \mathbb{P}\{-x + \log N + o(1) > W_{(N)}'\} \\
&= (1 - \exp(x - \log N + o(1)))^N \\
&= (1 - \exp(x + o(1))/N)^N \\
&\to \exp(-\exp x) \quad \text{as } N \to \infty.
\end{aligned}$$

That is, X_N has a limiting $\log W$ distribution. Furthermore, the weighted sum of independent variables on the right-hand side of (10) converges in $\mathscr{L}_2$, and hence almost surely to some random variable. Deduce that X_N converges to a $\log W$ variable. □

To establish the nonidentifiability of θ, we will construct a Y_θ independent of a $\log W$ variable such that $\log W =_D \theta(\log W + Y_\theta)$, for some fixed $0 < \theta < 1$. That is, there is no way to distinguish observations on W, a Weibull with shape parameter 1, and observations from the $[W\exp(Y_\theta)]^\theta$ scale mixture with shape parameter θ.

By the representation (10), we see that in order to demonstrate the lack of identification, we need to find a Y_θ, independent of an i.i.d. sequence $W_j' \sim \exp(1)$, such that

$$\left(\theta(\log W + Y_\theta) = -\theta\gamma - \sum_{j=1}^{\infty} (\theta W_j' - \theta)/j + \theta Y_\theta\right) =_D -\gamma - \sum_{j=1}^{\infty} (W_j' - 1)/j$$

for $0 < \theta < 1$. This suggests the choice

$$\theta Y_\theta = -\gamma(1-\theta) - \sum_{j=1}^{\infty} (B_j - 1 + \theta)/j,$$

where we need to find variables B_j independent of the W_j' such that $W_j' =_D \theta W_j' + B_j$. By comparing characteristic functions on the left- and right-hand sides of the last equality in distribution, we are forced to find a variable with the characteristic function

$$(1 - \theta it)/(1 - it) = \theta + (1-\theta)/(1-it) \quad \text{for all real } t.$$

Remarkably, such a variable does exist. If we let σ_0 denote a $\mathrm{Bin}(1, 1-\theta)$ variable independent of $W_0 \sim \exp(1)$, then the variable we seek is $\sigma_0 W_0$.

Let σ_j and W_j be sequences of variables that have the distribution of σ_0 and W_0, respectively. Furthermore, suppose that σ_j, W_j and W_j' are mutually independent. If we define Y_θ by

$$\theta Y_\theta = -\gamma(1-\theta) - \sum_{j=1}^{\infty} (\sigma_j W_j - 1 + \theta)/j \quad \text{for } 0 < \theta \le 1, \tag{11}$$

then we arrive at the equality $\log W =_D \theta(\log W + Y_\theta)$, where Y_θ is independent of $\log W$. This establishes the lack of identification in the Weibull mixture.

REMARK 12. Note that (11) is well defined because the sum of independent variables on the right-hand side converges in $\mathscr{L}_2$, and hence almost surely.

General conditions for identifiability. Return to the general problem of identifiability of θ from knowledge of the distribution for $\theta(X + Y)$. As the Weibull example shows, it might be possible to have distinct θ_1 and θ_2, for which

$$\theta_1(X + Y_1) =_D \theta_2(X + Y_2). \tag{13}$$

Equivalently,

$$\theta(X + Y_1) =_D X + Y_2, \tag{14}$$

where we may assume $\theta = \theta_1/\theta_2 \le 1$. Thereom 15 will show that the identification expressed by (13) cannot occur when $\theta < 1$ if the distributions for Y_1 and Y_2 are assumed to satisfy an exponential moment condition.

To express these conditions, we present the following method for measuring the tail behavior of a distribution. For each distribution G and each $0 < M \le \infty$, let

$$r_0(G, M) = \sup\{r \ge 0 : G\exp(\pm rY) < 1 + M\}.$$

Define $\mathscr{G}_M(r)$ to be the class of distributions G with $r_0(G, M) = r$ and let $\mathscr{G}_M(R+) = \bigcup_{r \geq R} \mathscr{G}_M(r)$.

THEOREM 15. *Suppose that* $0 < R = r_0(F, \infty) < \infty$. *Then the identification expressed by* (13) *is only possible when* $\theta_1 = \theta_2$ *and* $Y_1 =_D Y_2$ *if the distributions of* Y_1 *and* Y_2 *are required to belong to* $\mathscr{G}_M(R+)$ *for* $0 < M \leq \infty$.

Note that, in particular, Theorem 15 shows that the θ shape parameter in (1) is identified under the constraint that Y has a distribution in $\mathscr{G}_M(1+)$. This can be compared to the result in Heckman and Singer (1984), who show that θ is identified under the restriction $G \exp(-Y) < \infty$.

PROOF OF THEOREM 15. By the Tonelli–Fubini theorem and the tail behavior for X and Y_1,

$$\begin{aligned} \mathbb{P}\exp(\pm r\theta(X + Y_1)) &= \mathbb{P}(\exp(\pm r\theta X))\mathbb{P}(\exp(\pm r\theta Y_1)) \\ &< \infty \quad \text{for } 0 < r < R/\theta. \end{aligned} \tag{16}$$

The left-hand side of (16) must equal $\mathbb{P}\exp(\pm r(X + Y_2))$ by (14). Therefore, one more application of the Tonelli–Fubini theorem gives

$$\mathbb{P}(\exp(\pm rX))\mathbb{P}(\exp(\pm rY_2)) < \infty \quad \text{for } 0 < r < R/\theta.$$

The second factor on the left-hand side does not vanish. Therefore, if $\theta < 1$, we would arrive at the contradiction $\mathbb{P}(\exp(\pm rX)) < \infty$ for $R < r < R/\theta$. It follows that θ must equal 1.

Let $\hat{f}$ denote the Fourier transform for X, and let $\hat{g}_1$ and $\hat{g}_2$ denote the transforms for Y_1 and Y_2. With $\theta = 1$, the equality (14) implies that

$$\hat{f}(z)\hat{g}_1(z) = \hat{f}(z)\hat{g}_2(z) \quad \text{for } \{z : |\Im(z)| < R\}.$$

The analyticity of $\hat{f}$ guarantees that the transform must be nonzero in some open region within this strip (otherwise it would be 0 everywhere). Therefore, $\hat{g}_1$ must agree with $\hat{g}_2$ within this open region, and, hence, by analyticity the two transforms must agree along the real axis, yielding $Y_1 =_D Y_2$. □

4. Rates of estimation for square summable exponentials. Theorem 15 presented sufficient conditions that ensure identification for the location mixture model (4). In this section we consider how these constraints play a role in the estimation for the unknown scale parameter in (4). The crux of the theory revolves around the relationship between the severity of moment constraints imposed on the mixing Y (to ensure identifiability) and the manner in which these constraints affect the Hellinger distance between mixture models with differing parameters.

To motivate the discussion, first consider the problem of estimation for the θ parameter in the transformed Weibull mixture (3). Suppose then that $\mathscr{P}$ is the identified class of mixtures of the form (3) whose Y have distributions in $\mathscr{G}_M((1 + d)+)$ for some $0 < M \leq \infty$ and $d \geq 0$. Let us consider how well θ can be estimated.

In order to derive a sharp lower rate of estimation for θ, we will construct a Y_θ with distribution in $\mathscr{G}_M((1+d)+)$ that makes discrimination between the mixture $\theta(\log W + Y_\theta)$ and the $\log W$ variable as difficult as possible. Because the Y_θ presented in (11) represents the worst one-dimensional case for the unconstrained problem, the strategy we use will be to modify this variable slightly so as to satisfy the necessary moment constraints.

In order to determine how we might modify (11), let us calculate the tail behavior of its distribution. The contribution from each j-term in the sum (11) to $\mathbb{P}\exp(rY_\theta)$ equals

$$\exp\left(-\frac{(1-\theta)r}{j\theta}\right)\left(\theta + \frac{1-\theta}{1-r/(j\theta)}\right) \quad \text{for } \frac{r}{j} < \theta < 1.$$

If k is the largest integer less than or equal to $1+d$, then it is the first k terms in the sum of (11) that are problematic. For even when θ is near 1, their contribution to $\mathbb{P}\exp(rY_\theta)$ becomes unbounded as r approaches $j \le 1+d$. As we shall see, these are the only terms that cause any difficulties. Consequently, one method for satisfying the moment constraint would be to replace the k problematic terms by truncated versions

$$(\sigma_j W_j^* - 1 + \theta)/j \quad \text{for } j = 1, \ldots, k, \tag{17}$$

where $W_j^* \sim C_\tau \exp(-w)$, $0 < w < \tau$, are truncated exponential variables assumed to be independent of σ_j, and $C_\tau = 1/(1-\exp(-\tau))$ for $0 < \tau < \infty$.

Our strategy, then, will be to replace a finite number of terms in (11) with the modified variables (17). By allowing $\tau \to \infty$ as rapidly as possible in order to comply with the moment constraints, we will be able to construct a mixture $\theta(\log W + Y_\theta)$ in $\mathscr{P}$ that is difficult to distinguish from the $\log W$ variable. This method will generate our rates of estimation.

The same strategy can be extended to apply to the general mixture (4) for the class of X variables that can be written as the sum of independent $W_j' \sim \exp(1)$ variables:

$$X = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j (W_j' - 1), \tag{18}$$

where $0 > |\alpha_1| \ge |\alpha_2| \ge \cdots$, $|\alpha_0| < \infty$ and $\sum_{j=1}^{\infty} \alpha_j^2 < \infty$. For the general case, lower rates will be established using a nuisance mixing variable defined as follows. For $k \ge 0$ and $0 < \theta \le 1$, define $Y_{k,\theta}$ by

$$\theta Y_{k,\theta} = \alpha_0(1-\theta) + \sum_{j=1}^{k} \alpha_j(\sigma_j W_j^* - 1 + \theta) + \sum_{j=k+1}^{\infty} \alpha_j(\sigma_j W_j - 1 + \theta), \tag{19}$$

where σ_j, W_j and W_j^* are assumed to be mutually independent.

THEOREM 20. *Let $\mathscr{P}$ be the (identified) class of mixtures* (4) *with $X \sim F$ of the form* (18) *and Y with distribution in $\mathscr{G}_M(R_d+)$ for $0 < M \le \infty$ and $R_d = (1+d)r_0(F,\infty)$. Suppose $\theta_0(X + Y_0)$ is a mixture in $\mathscr{P}$, where Y_0 has distribution in $\mathscr{G}_M(R_d'+)$ for $R_d' > R_d$. Then there exists a family $\mathscr{P}_n \subseteq \mathscr{P}$*

that converges uniformly in $\mathscr{L}_1$ *to* $\theta_0(X + Y_0)$ *such that* θ *can be estimated at a rate no faster than* $O_p(1/\log n)$ *over* $\mathscr{P}_n$ *when* $d = 0$ *and* $O_p(n^{-d/(2d+1)})$ *when* $d > 0$.

The theorem applies to the Weibull semiparametric mixture (1). In particular, we find that even though θ is identified when $d = 0$, it still cannot be estimated at better than an $O_p(1/\log n)$ rate. Interestingly, this shows that the first moment constraint assumed by Heckman and Singer (1984), although sufficient to ensure consistency of their nonparametric maximum likelihood estimator, is not strict enough to guarantee a polynomial rate of estimation. Polynomial rates can only be achieved under the more stringent moment constraint when $d > 0$. In particular, the theorem shows that rates of estimation approach $O_p(1/\sqrt{n})$ only as $d \to \infty$.

PROOF OF THEOREM 20. To establish the rates, we will use (19) to construct a mixture in $\mathscr{P}$ difficult to distinguish from the $\theta_0(X + Y_0)$ mixture. Because we can always rescale θ by θ_0, we can assume without loss of generality that $\theta_0 = 1$. Furthermore, we could always add an independent Y_0 term to the right-hand side of (19) in constructing our mixture. Because $R'_d > R_d$, the distribution for Y_0/θ will eventually lie in $\mathscr{G}_M(R_d+)$ for θ values close enough to 1. Therefore, the contribution from Y_0 to our construction would not undermine our efforts to ensure that (19) satisfied the necessary moment constraints. Thus, without loss of generality, we can assume that $Y_0 = 0$ and $\theta_0 = 1$.

Let k be the first integer so that $R_d = (1 + d)/|\alpha_1| < 1/|\alpha_{k+1}|$. The asserted lower rates will follow from a Hellinger distance calculation between X and the mixture $\theta(X + Y_{k,\theta})$, where $Y_{k,\theta}$ is defined by (19) with

$$\tau = \begin{cases} \theta/[(1-\theta)\rho], & \text{for } d = 0, \\ -\theta(\log(1-\theta) + \rho)/(d + 1 - \theta), & \text{for } d > 0, \end{cases} \tag{21}$$

and $\rho > 0$ is yet to be specified (we assume that $0 < \theta \leq 1$ is close enough to 1 to ensure that $\tau > 0$).

Of course, we first need to verify that $\theta(X + Y_{k,\theta})$ belongs to $\mathscr{P}$ for the truncation level specified in (21). To do so, it suffices to show that $\mathbb{P}\exp(\pm R_d Y_{k,\theta}) < 1 + M$.

From the inequality $1 + x \leq \exp(x)$, deduce that

$$\begin{aligned} &\mathbb{P}\exp\big(r(\sigma_j W_j - 1 + \theta)\big) \\ &\quad = \exp(-r(1-\theta))\left(1 + \frac{(1-\theta)r}{1-r}\right) \\ &\quad \leq \exp\left(\frac{(1-\theta)r^2}{|1-|r||}\right) \quad \text{for } r < 1 \text{ and } 0 \leq \theta \leq 1. \end{aligned} \tag{22}$$

Our choice for k ensures that $R_d|\alpha_j|$ is strictly less than 1 uniformly for $j \geq k+1$. Use (22) over the $j \geq k+1$ terms in (19) to show that $\mathbb{P}\exp(\pm R_d Y_{k,\theta})$ is smaller than

$$\exp\left(\frac{(1-\theta)R_d(|\alpha_0| + k|\alpha_1|)}{\theta}\right)\left(\prod_{j=1}^{k}\mathbb{P}\exp\left(\frac{R_d|\alpha_j|\sigma_j W_j^*}{\theta}\right)\right)$$
$$\times \exp\left(\sum_{j=k+1}^{\infty}\frac{(1-\theta)R_d^2\alpha_j^2/\theta^2}{|1 - R_d|\alpha_j|/\theta|}\right).$$

The denominator in the summation of the above expression remains strictly bounded away from 0 for θ close to 1. This and the square summability of α_j shows that the expression is bounded by

$$(1 + o(1))\left(\mathbb{P}\exp((1+d)\sigma_1 W_1^*/\theta)\right)^k,$$

where the $o(1)$ term is uniform as $\theta \uparrow 1$. Meanwhile,

$$\mathbb{P}\exp\left(\frac{(1+d)\sigma_1 W_1^*}{\theta}\right) = \theta + \frac{\theta(1-\theta)C_\tau}{(d+1-\theta)}\left[\exp\left(\frac{(d+1-\theta)\tau}{\theta}\right) - 1\right],$$

which by (21) is less than or equal to

$$\begin{cases} 1 + C_\tau(\exp(1/\rho) - 1) = 1 + o(1), & \text{for } d = 0, \\ 1 + C_\tau\exp(-\rho)/(d+1-\theta) = 1 + o(1), & \text{for } d > 0, \end{cases}$$

where the $o(1)$ variable is uniform as $\theta \uparrow 1$ and $\rho \uparrow \infty$. This proves that $Y_{k,\theta}$ has distribution in $\mathscr{G}_M(R+)$ for a suitably large enough ρ and θ close enough to 1.

Now to establish the rate. The same argument given in Section 3 implies that $X =_D \theta(X + Y_{0,\theta})$. Therefore, $k-1$ applications of the triangle inequality gives

$$\begin{aligned} H(X, \theta(X + Y_{k,\theta})) &= H(\theta(X + Y_{0,\theta}), \theta(X + Y_{k,\theta})) \\ &\leq \sum_{j=1}^{k} H\big(\theta(X + Y_{j-1,\theta}), \theta(X + Y_{j,\theta})\big). \end{aligned} \tag{23}$$

The Hellinger distance satisfies $H(V + V_1, V + V_2) \leq H(V_1, V_2)$ for any random variable V independent of any V_1 and V_2. We can assume that $\{\sigma_j, W_j, W_j', W_j^*\colon j \geq 1\}$ are mutually independent. Therefore, because the Hellinger distance is invariant under a change of scale and location, we can bound the right-hand side of (23) by

$$\sum_{j=1}^{k} H\big(\theta W_j' + \sigma_j W_j, \theta W_j' + \sigma_j W_j^*\big) = kH(W_1', \theta W_1' + \sigma_1 W_1^*), \tag{24}$$

where the last equality follows from the identity $W_j' =_D \theta W_j' + \sigma_j W_j$. Let m denote the density for W_1. Then the same identity shows that

$$m = m\left(\frac{\cdot}{\theta}\right) + (1-\theta)\left[\frac{1}{\theta}m\left(\frac{\cdot}{\theta}\right) * m\right].$$

Furthermore, if m_θ denotes the density for the truncated exponential W_1^*, then $\theta W_1' + \sigma_1 W_1^*$ has density

$$m\left(\frac{\cdot}{\theta}\right) + (1-\theta)\left[\frac{1}{\theta}m\left(\frac{\cdot}{\theta}\right) * m_\theta\right] = m + (1-\theta)\Delta_\theta,$$

where

$$\Delta_\theta = \frac{1}{\theta}m\left(\frac{\cdot}{\theta}\right) * [m_\theta - m].$$

Hence, deduce from (24) that

$$\begin{aligned} H(X, \theta(X + Y_{k,\theta})) &\le k\sqrt{\int \left(\sqrt{m} - \sqrt{m + (1-\theta)\Delta_\theta}\right)^2} \\ &\le k(1-\theta)\sqrt{\int \frac{\Delta_\theta^2}{m}}, \end{aligned} \tag{25}$$

where the last inequality follows from $(\sqrt{a} - \sqrt{b})^2 \le (a-b)^2/a$ for any $a, b \ge 0$.

Let $1\{\cdot\}$ denote the indicator function. To determine the contribution from the integral on the right-hand side of (25), use the bound

$$\theta|\Delta_\theta(w)| \le \exp(-w)\left[w(C_\tau - 1)1\{w < \tau\} + (\tau(C_\tau - 2) + w)1\{w \ge \tau\}\right],$$

to show that

$$\begin{aligned} \theta^2 \int \frac{\Delta_\theta(w)^2}{m(w)}\,dw &\le (C_\tau - 1)^2 \int_0^\tau w^2 \exp(-w)\,dw \\ &\quad + \exp(-\tau)\int_0^{+\infty} (\tau(C_\tau - 1) + w)^2 \exp(-w)\,dw \\ &= O(\exp(-2\tau)) + O(\exp(-\tau)), \end{aligned}$$

because $C_\tau - 1 = O(\exp(-\tau))$. Therefore,

$$H(X, \theta(X + Y_{k,\theta})) \le' O((1-\theta)\exp(-\tau/2)). \tag{26}$$

The asserted rates will now follow by allowing θ to depend on the sample size n. Let $\tau = \theta_n/[(1-\theta_n)\rho] = \log n$ when $d = 0$, and let $\theta_n = 1 - \varepsilon n^{-d/(2d+1)}$ when $d > 0$ [note that this defines τ by (21)]. Therefore, from (26),

$$H\left(X, \theta_n(X + Y_{k,\theta_n})\right) \le \begin{cases} O(n^{-1/2}/\log n), & \text{for } d = 0, \\ \varepsilon' n^{-1/2}, & \text{for } d > 0, \end{cases} \tag{27}$$

where $\varepsilon' > 0$ can be made arbitrarily small by varying ε.

Let $\mathscr{P}_n \subseteq \mathscr{P}$ be the family of mixtures X and $\theta(X + Y_{k,\theta})$, where $\theta_n \leq \theta < 1$. Then we have exhibited mixture models in $\mathscr{P}_n$ whose structural parameters are separated by $O(1/\log n)$ when $d = 0$ and $O(n^{-d/(2d+1)})$ when $d > 0$, and whose Hellinger distance is a small multiple of $n^{-1/2}$. Furthermore, because the Hellinger distance (27) bounds the $\mathscr{L}_1$-distance, $\mathscr{P}_n$ converges uniformly in $\mathscr{L}_1$ to X. This establishes the asserted rates. □

Acknowledgments. This paper evolved from ideas and results contained in my Ph.D. thesis. A large part of this evolution was sparked from the countless discussions and exchange of ideas between myself and my dissertation advisor, David Pollard. In particular, he contributed some of the ideas related to the series expansion of the log exponential. I would also like to thank the anonymous Associate Editor who suggested the moment bounds used in the proof of Theorem 20.

REFERENCES

DONOHO, D. L. and LIU, R. C. (1987). Geometrizing rates of convergence. I. Technical report, Dept. Statistics, Univ. California, Berkeley.

DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence. II. *Ann. Statist.* **19** 633–667.

HECKMAN, J. and SINGER, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52** 271–320.

HONORÉ, B. E. (1990). Simple estimation of a duration model with unobserved heterogeneity. *Econometrica* **58** 453–473.

HONORÉ, B. E. (1994). A note on the rate of convergence of estimators of mixtures of Weibulls. Unpublished manuscript.

ISHWARAN, H. (1996). Uniform rates of estimation in the semiparametric Weibull mixture model. *Ann. Statist.* **24** 1572–1585.

JEWELL, N. P. (1982). Mixtures of exponential distributions. *Ann. Statist.* **10** 479–484.

LE CAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53.

UNIVERSITY OF OTTAWA
DEPARTMENT OF MATHEMATICS AND STATISTICS
P.O. BOX 450, STN A
OTTAWA, ONTARIO
CANADA K1N 6N5
E-MAIL: ishwaran@expresso. mathstat.uottawa.ca