Uniform Rates of Estimation in the Semiparametric Weibull Mixture Model

Hemant Ishwaran

Annals of Statistics, Volume 24, Issue 4 (Aug., 1996), 1572-1585.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199608%2924%3A4%3C1572%3AUROEIT%3E2.0.CO%3B2-K

Annals of Statistics

The Annals of Statistics
1996, Vol. 24, No. 4, 1572–1585

UNIFORM RATES OF ESTIMATION IN THE SEMIPARAMETRIC WEIBULL MIXTURE MODEL

BY HEMANT ISHWARAN

University of Ottawa

This paper presents a uniform estimator for a finite-dimensional parameter in the semiparametric Weibull mixture model. The rates achieved by the estimator hold uniformly over shrinking sequences of models much more general than traditional sequences that are required to satisfy a Hellinger differentiable property. We show that these rates are optimal in a class of identified models constrained by a moment condition on the nonparametric mixing distribution.

1. Introduction. The intention of this paper is to study the semiparametric Weibull mixture model described in Heckman and Singer (1984), with the intent of presenting an optimal uniform estimator for a finite-dimensional parameter in the model. The paper explores the connection between identification constraints imposed on the nonparametric component in the model (the unknown mixing distribution) and the manner in which these constraints affect achievable uniform rates of estimation.

The model that we study will be assumed to come from duration data (T, Z), where T is the observed positive duration time and Z is an observed k-dimensional vector of covariates. We also assume that potential heterogeneity may enter the data via an unobserved heterogeneity variable Y, with unknown distribution G, that is assumed to be independent of Z. Write h for the unknown density of Z taken with respect to a σ-finite measure ν. We assume, as in Heckman and Singer (1984), that (T, Z) has the semiparametric Weibull mixture density

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

where (β, θ) denotes the $(k+1)$-dimensional structural parameter and G denotes the unknown nuisance mixing distribution. The shape parameter θ is assumed to be strictly positive, while the covariate parameter β lies in $\mathbb{R}^k$.

Several authors have studied the Weibull regression model with unobserved heterogeneity. Heckman and Singer (1984) verified the consistency of a semiparametric maximum likelihood estimator for (β, θ, G) under a first exponential moment constraint to G. Honoré (1990) proposed a class of

Received October 1994; revised December, 1995.

AMS 1991 *subject classifications.* Primary 62G05; secondary 62G20, 62P20.

Key words and phrases. Weibull semiparametric mixture, uniform rates of estimation, criterion function.

estimators based on order statistics that achieves an $O_p(n^{-s}/\log n)$ rate of estimation for θ for each $0 < s < 1/3$. The result was based on the assumption that no covariates are present in the model and that the mixing distribution has a finite second exponential moment (using our parameterization). Preliminary work by Honoré (1994) suggests that his class of estimators can be extended to achieve an $O_p(n^{-s}/\log n)$ rate under a $(1 + d)$th moment constraint for each $0 < s < d/(2d + 1)$, where $d \geq 1$ is any integer. Recently, Ishwaran (1996) showed (as a special case) that if G satisfies a $(1 + d)$th exponential moment constraint, then it is not possible to estimate θ at a uniform rate faster than $O_p(n^{-d/(2d+1)})$.

The contribution of this paper to the study of the Weibull mixture (1) will be to present a uniform estimator for θ that achieves the lower rate proposed in Ishwaran (1996) for the case when $0 < d \leq 1$. This will then establish $O_p(n^{-d/(2d+1)})$ as the optimal rate for estimating θ in a class of Weibull mixtures with covariates. Moreover, this would also indicate that $O_p(n^{-d/(2d+1)})$ is the optimal rate for estimating (β, θ) in this class, because constructing an estimator for β is relatively easy once we have an estimator for θ.

To see why this is the case, suppose that $W \sim \exp(1)$ and $T = [W \exp(\beta'Z + Y)]^{1/\theta}$, where W, Y, Z are assumed to be mutually independent. Then (T, Z) has the (conditional) semiparametric Weibull density (1). If we transform T by a log, then

$$X = \log T = \frac{1}{\theta}(\log W + \beta'Z + Y) \tag{2}$$

describes a regression problem with unknown regression parameter $\theta^{-1}\beta$, observed covariate Z and an unobserved smooth error. Constructing an $O_p(n^{-1/2})$ estimator for $\theta^{-1}\beta$ is relatively straightforward using semiparametric regression methods [see Bickel (1982) for a study of the general regression problem with unknown error]. Now to estimate the covariate parameter β, we only need to multiply a $\theta^{-1}\beta$ estimate with an estimate for θ.

Section 3 describes a method for constructing a uniform estimator for θ based upon a sample of n independent transformed observations $X_1, X_2, \ldots, X_n$ as in (2). Transforming the data will allow us to take advantage of X being a convolution. The key idea behind our method is that, when G is constrained by a moment condition, the density for (X, Z),

$$\begin{aligned} f(x, z|\beta, \theta, G) &= \theta \exp(\theta x) h(z) \exp(-\beta'z) \\ &\quad \times \int \exp(-y - \exp(\theta x - \beta'z - y))\, dG(y), \end{aligned} \tag{3}$$

can be approximated for large negative x as

$$\theta \exp(\theta x) h(z) \exp(-\beta'z) \int \exp(-y)\, dG(y). \tag{4}$$

Therefore, by smoothing the data with an appropriate kernel, the tail behavior of the smoothed data will be driven, to a first approximation, by the behavior of (4). A comparison of the smoothed data to its expected behavior under the approximation (4) provides a method for estimating θ. Section 3 presents an estimator for θ based on this argument. The main result of the paper is contained in the rate result found in Theorem 22 of Section 4. There we show that our estimator for θ achieves the optimal $O_p(n^{-d/(2d+1)})$ rate when G is constrained by a $(1+d)$th exponential moment condition, where $0 < d \leq 1$.

The rates of estimation given in this paper are of a uniform nature. By this we mean the following. Suppose $\mathscr{P} = \{\mathbb{P}_\gamma\colon \gamma \in \Upsilon\}$ is a class of transformed Weibull mixtures, where Υ is some parameter set of (β, θ, G) values which indexes our model space. Then, a uniform estimator for θ is an estimator that does uniformly well over different sampling schemes from $\mathscr{P}$.

To make this more precise, write $\mathbb{P}_\gamma$ for the distribution of the mixture (X, Z) with parameter $\gamma = (\beta, \theta, G) \in \Upsilon$. Let θ be the functional $\theta\colon \Upsilon \to \mathbb{R}_+$ which maps $\gamma \in \Upsilon$ onto its shape parameter $\theta(\gamma) = \theta$. Notice that we are using θ to denote both the parameter value and the functional (although there is some slight ambiguity in doing so, it will greatly simplify our notation). By uniform rates, we mean the following.

DEFINITION 5. Let Υ_n be a sequence of families in Υ and let δ_n be a decreasing positive sequence. Estimators $\hat{\theta}_n$ for θ are said to have a uniform $O_p(\delta_n)$ rate of convergence over Υ_n if for each $\varepsilon > 0$ there exists a finite constant κ_ε such that

$$\limsup_{n\to\infty} \sup_{\gamma\in\Upsilon_n} \mathbb{P}_\gamma^n\left\{\left|\hat{\theta}_n - \theta(\gamma)\right| \geq \kappa_\varepsilon \delta_n\right\} \leq \varepsilon,$$

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

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

2. Identifiability. Ishwaran (1996) shows by an explicit construction that the semiparametric mixture model (3) [and consequently (1)] is unidentified without constraints. The author describes constraints for Y which ensure that the mixture

$$X = \frac{1}{\theta}(\log W + Y) \tag{6}$$

is identified. More precisely, for each $0 < M < \infty$ define

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

to be a measure for the tail behavior of a distribution G. Define $\mathcal{G}_M(r)$ to be the class of distributions G with $r_0(G, M) = r$ and let $\mathcal{G}_M(R+) = \bigcup_{r \geq R} \mathcal{G}_M(r)$. Then, Ishwaran (1996) shows that the model (6) is identified under the constraint that $G \in \mathcal{G}_M(1+)$ for each $0 < M < \infty$.

With some minimal assumptions on the distribution for Z, we can ensure that the Weibull mixture model with covariates is identified. Write $\mathbb{P}_Z$ for the distribution of Z. We say that $\mathbb{P}_Z$ is nondegenerate if $\mathbb{P}_Z\{\beta'Z = c\} = 1$ implies that $\beta = 0$ and $c = 0$ (for example, nondegeneracy implies that the support for Z contains a basis for $\mathbb{R}^k$ and elements which have at least two different values in each vector coordinate).

THEOREM 7 [Compare with Heckman and Singer (1984)]. *Suppose that* $d > 0$ *and* $0 < M < \infty$. *If* $\mathbb{P}_Z$ *is nondegenerate, then the identification expressed by*

$$\left(\frac{1}{\theta_1}(\log W + \beta_1'Z + Y_1), Z\right) =_{\mathcal{D}} \left(\frac{1}{\theta_2}(\log W + \beta_2'Z + Y_2), Z\right) \tag{8}$$

for each pair $(\beta_j, \theta_j, G_j) \in \mathbb{R}^k \otimes (0, \infty) \otimes \mathcal{G}_M((1 + d)+)$, *is only possible if* $(\beta_1, \theta_1) = (\beta_2, \theta_2)$ *and* $Y_1 =_{\mathcal{D}} Y_2$, *where* Y_j *denotes the variable with distribution* G_j.

PROOF. The equality between the two mixtures (8) implies that

$$\frac{1}{\theta_1}(\log W + \beta_1'z + Y_1) =_{\mathcal{D}} \frac{1}{\theta_2}(\log W + \beta_2'z + Y_2), \quad \text{for a.a. } z[\mathbb{P}_Z]. \tag{9}$$

Suppose that z_0 is a value in the support for Z where (9) holds. Let $Y_j^* = \beta_j'z_0 + Y_j$ with distribution G_j^*, for $j = 1, 2$. Because $G_j^* \in \mathcal{G}_{M^*}(1+)$ for a large enough $0 < M^* < \infty$, Theorem 15 of Ishwaran (1996) shows (9) implies that $\theta_1 = \theta_2$ and $Y_1^* =_{\mathcal{D}} Y_2^*$. Therefore,

$$\log W + \beta_1'Z + Y_1 =_{\mathcal{D}} \log W + \beta_2'Z + Y_1 + (\beta_1 - \beta_2)'z_0. \tag{10}$$

If we write $\hat{g}$ for the Fourier transform of $\log W + Y_1$ and $\hat{h}_{\beta_j}$ for the transform of $\beta_j'Z$, the equality (10) becomes

$$\hat{g}(s)\hat{h}_{\beta_1}(s) = \hat{g}(s)\hat{h}_{\beta_2}(s)\exp(is(\beta_1 - \beta_2)'z_0) \quad \text{for all real } s.$$

The tail behavior of the distribution for $\log W + Y_1$ ensures that its transform must be analytic in the horizontal strip containing complex values with imaginary parts less than 1. Therefore, $\hat{g}$ must be nonzero (otherwise it would be zero throughout its region of analyticity). Dividing throughout by $\hat{g}$, deduce that

$$\beta_1'Z =_{\mathcal{D}} \beta_2'Z + (\beta_1 - \beta_2)'z_0.$$

Indeed, our argument shows that we can replace z_0 on the right-hand side by a.a. values of $z[\mathbb{P}_Z]$. The nondegeneracy of $\mathbb{P}_Z$ now gives $\beta_1 = \beta_2$, and consequently that $Y_1 =_{\mathscr{D}} Y_2$. □

3. Construction of an estimator. Even without the presence of a covariate, Ishwaran (1996) shows that it is not possible to estimate the θ parameter in the Weibull mixture at a uniform rate faster than $O_p(n^{-d/(2d+1)})$. This result holds over the class of identified mixtures whose mixing distributions are members of $\mathscr{G}_M((1+d)+)$, where $d > 0$ and $0 < M < \infty$. The remainder of this paper will be devoted to showing that this rate is in fact the optimal uniform rate for θ in (3) for fixed values $0 < d \leq 1$ and $0 < M < \infty$. Because from here on we will only consider d and M values fixed within this range, we will hereafter suppress their use, whenever possible, unless there is ambiguity.

Write Υ for the index set $\mathscr{B} \otimes \Theta \otimes \mathscr{G}_M((1+d)+)$, where $\mathscr{B}$ and Θ denote the parameter spaces for β and θ, respectively. Let λ be the functional $\lambda\colon \Upsilon \to \mathbb{R}_+^2$ which maps each $\gamma = (\beta, \theta, G) \in \Upsilon$ onto $\lambda(\gamma) = (\theta, \phi)$, where

$$\phi = (G\exp(-Y))(\mathbb{P}_Z \exp(-\beta' Z)).$$

For notational convenience we will write $\lambda(\Upsilon)$ for the range of λ. Also, following the convention used with the θ functional in Section 1, we will use $\lambda = (\theta, \phi)$ to denote both the parameter value and the functional itself: $\lambda = \lambda(\gamma)$.

The identification argument of the previous section and the lower rate calculation given in Ishwaran (1996) implicitly rely on the left exponential tail behavior for the density of a log exponential variable. It is this left tail behavior that we will exploit in constructing an estimator for θ in (3). Let K be a kernel density and write $m_\theta(x) = \theta \exp(\theta x)\{x \leq 0\}$ to denote the density for a negative exp(1) random variable. Then our optimal estimator for θ will be based on the value of $\lambda = (\theta, \phi)$ in $\lambda(\Upsilon)$ which minimizes, or nearly minimizes, the following criterion function

$$\Gamma_n(\lambda, C_n) = \int_{-\infty}^{-C_n} \left(\frac{1}{n} \sum_{i=1}^{n} K(u - X_i) - \phi[K * m_\theta](u) \right)^2 du. \tag{11}$$

Here $C_n \to \infty$ is a positive sequence yet to be specified. Notice that the nuisance distribution function G is eliminated from the optimization problem, except for the estimation of the scalar nuisance parameter ϕ.

The heuristic which leads to this method is as follows. Suppose that we are sampling under the model $\mathbb{P}_{\gamma_n}$ for (X, Z), where $\gamma_n = (\beta_n, \theta_n, G_n)$ is a parameter in Υ. If we let $h_{\beta_n}(z) = h(z)\exp(-\beta_n' z)$, then the density (3) for $\mathbb{P}_{\gamma_n}$ can be expressed as

$$f(x, z|\gamma_n) = f_1(x, z|\gamma_n) + f_2(x, z|\gamma_n), \tag{12}$$

where

$$f_1(x, z|\gamma_n) = [G_n \exp(-Y)] m_{\theta_n}(x) h_{\beta_n}(z)$$

and

$$f_2(x, z|\gamma_n) = \theta_n \exp(\theta_n x) h_{\beta_n}(z) \times \int \exp(-y)\left[\exp(-\exp(\theta_n x - \beta_n' z - y)) - \{x \le 0\}\right] dG_n(y).$$

If $w(x, y, z) = \exp(\theta_n x - \beta_n' z - y)$, then we can express $f_2(x, z|\gamma_n)$ for $x \le 0$, as

$$m_{\theta_n}(x) h_{\beta_n}(z) \exp(d(\theta_n x - \beta_n' z)) \times \int \left[\frac{\exp(-w(x, y, z)) - 1}{w(x, y, z)^d}\right] \exp(-(1 + d)y)\, dG_n(y). \tag{13}$$

The moment constraint imposed on G_n and the inequality $|\exp(-w) - 1|/w^d \le 1$, for $w \ge 0$, imply that (13) is bounded by

$$(1 + M) m_{\theta_n}(x) h_{\beta_n}(z) \exp(d(\theta_n x - \beta_n' z)). \tag{14}$$

Consequently, for $x \le 0$, write

$$f(x, z|\gamma_n) = f_1(x, z|\gamma_n)\left[1 + O(\exp(d(\theta_n x - \beta_n' z)))\right]$$

to indicate that it is the f_1 contribution which drives the left tail behavior of f for large negative x. Notice that the contribution from the smaller order term depends directly upon the rate of decrease imposed on the tails of the distribution G_n.

Forgoing exact details for the moment, assume that the kernel K is some smooth density over the real line with rapidly decreasing tails. Then, approximately, the average of the smoothed data should be

$$\frac{1}{n} \sum_{i=1}^{n} K(u - X_i) \approx \mathbb{P}_{\gamma_n} K(u - X). \tag{15}$$

If h_{β_n} is ν-integrable, then the expression for the density (12) shows that

$$\mathbb{P}_{\gamma_n} K(u - X) = \phi_n\left[K * m_{\theta_n}\right](u) + \Delta_{\gamma_n}(u), \tag{16}$$

where $\phi_n = (G_n \exp(-Y))(\mathbb{P}_Z \exp(-\beta_n' Z))$ and

$$\Delta_{\gamma_n}(u) = \int\int K(u - x) f_2(x, z|\gamma_n)\, dx\, d\nu(z).$$

The behavior of Δ_{γ_n} depends upon the behavior of f_2, while the behavior of the convolution $\phi_n[K * m_{\theta_n}]$ depends upon f_1, which as we have seen is the dominant of the two terms for large negative values. Therefore, if the tails for K decrease rapidly enough, we expect Δ_{γ_n} to be small in comparison to the convolution. Thus, on average, we expect the smoothed data on the left-hand side of (15) to be close to the convolution when u is negative and large. This leads to our method of estimation based on (11).

Although the results in the remainder of the paper can be proved using any density kernel (with tails decreasing fast enough), many of the proofs are

greatly simplified by working with a prespecified K. In particular, choosing K to be a uniform(0, 1) kernel is especially appealing. In this case

$$[K * m_{\theta_n}](u) = \exp(\theta_n u)(1 - \exp(-\theta_n)) \quad \text{when } u \leq 0, \tag{17}$$

showing that the tail behavior for the convolution is driven solely by the parameter of interest θ_n. For simplicity we will assume that K is the uniform(0, 1) density.

To ensure that our rates will hold uniformly over a given family of models, we will need to control the behavior of the corresponding θ parameters. A family of models will be said to be *regular* if the following is true.

DEFINITION 18. Families $\Upsilon_n \subseteq \Upsilon$ are said to be regular if $\theta(\Upsilon_n)$ converges to an interior point of Θ such that for any pair $\theta_1, \theta_2 \in \theta(\Upsilon_n)$,

$$|\theta_1 - \theta_2| \leq A(\log n)^{-1},$$

for some fixed $A < \infty$.

Note the generality of the shrinking sequences that will be implicit in our uniform rates. In particular, our rates will not be restricted to sequences that are required to satisfy a Hellinger differentiable property as is the traditional practice in semiparametric estimation [see for example Begun, Hall, Huang and Wellner (1983)].

Two further assumptions that are needed to ensure uniformity will be that the parameter space for β and θ is compact and that the density for Z has tails which decrease rapidly enough. Hereafter, the following will be assumed.

ASSUMPTION 19. (i) $\mathscr{B} \otimes \Theta = \{(\beta, \theta): |\beta| \leq b_0,\ t_0 \leq \theta \leq t_1\}$, where $0 < b_0 < \infty$ and $0 < t_0 < t_1 < \infty$ are known fixed values. (ii) $\mathbb{P}_Z \exp((1 + d)b_0|Z|) < \infty$.

One consequence of Assumption (19)(i) is that the parameter space $\lambda(\Upsilon)$ is bounded and strictly finite (see Lemma 20). A straightforward argument using the dominated convergence theorem and identity (17) shows that (11) is continuous in λ. Therefore, an important consequence to Lemma 20 is that a global minimizer of (11) exists and lies in $\lambda(\Upsilon)$.

LEMMA 20. $\lambda(\Upsilon) = [t_0, t_1] \otimes [p_0, p_1]$, *where* $0 < t_0 < t_1 < \infty$ *and* $0 < p_0 < p_1 < \infty$.

The proof of Lemma 20 is given in the Appendix.

4. Uniform rates of estimation. In order to construct our optimal uniform estimator, we will need the existence of a preliminary uniform estimator for θ in order to fine-tune the cutoff sequence used in our criterion function. More precisely, if $\tilde{\theta}_n$ is a preliminary $o_p(1/\log n)$ uniform estimator,

then our optimal estimator is obtained from nearly minimizing $\Gamma_n(\lambda, \tilde{C}_n)$, where $\tilde{C}_n = \tilde{C}_n(\tilde{\theta}_n)$ is a random cutoff sequence based on $\tilde{\theta}_n$.

Using a standard comparison argument, Ishwaran (1995) establishes the existence of a uniform $o_p(1)$ estimator for θ. Note that by uniform consistency, or for that matter uniform probability, we always mean uniformity over some regular sequence of families Υ_n. Thus, for example, a uniform o_p statement such as

$$\zeta_n = o_p(1) \quad \text{uniformly over } \Upsilon_n,$$

will mean that for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\gamma\in\Upsilon_n} \mathbb{P}_\gamma^n\{|\zeta_n| \geq \varepsilon\} = 0.$$

Although the result in Ishwaran (1995) only asserts the existence of an $o_p(1)$ estimator, it is still possible to use this estimator to construct the $o_p(1/\log n)$ estimator needed in our proof of optimality (in fact using a consistent estimator we can construct a near optimal estimator for θ). The technical details of this construction, along with the proof of consistency, can be found in the same paper. Those details will be omitted here, but for completeness the consistency result is stated in the following theorem. It should be noted that the key condition in the theorem is to choose the cutoff sequence $\{C_n\}$ so that it increases at a rate no faster than the log of a power of n.

THEOREM 21. *Let Υ_n be a regular sequence of families. Suppose that $C_n \to \infty$ with uniform probability tending to 1 and for some $0 < s < 1$,*

$$C_n \leq \frac{\log(n^s)}{\inf\{\theta\colon \theta(\Upsilon_n)\}} + o_p(1) \quad \textit{uniformly over } \Upsilon_n.$$

If the estimator $\hat{\lambda}_n = (\hat{\theta}_n, \hat{\phi}_n) \in \lambda(\Upsilon)$ satisfies

$$\Gamma_n\big(\hat{\lambda}_n, C_n\big) \leq o_p(n^{-2s}) + \inf_{\lambda\in\lambda(\Upsilon)} \Gamma_n(\lambda, C_n) \quad \textit{uniformly over } \Upsilon_n,$$

then for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\gamma\in\Upsilon_n} \mathbb{P}_\gamma^n\big\{|\hat{\theta}_n - \theta(\gamma)| \geq \varepsilon\big\} = 0.$$

Notice that the proposed estimator in Theorem 21 is not required to uniquely minimize $\Gamma_n(\lambda, C_n)$. Allowing for a near minimizer is possible without requiring any additional assumptions, thus permitting a slightly more general theorem. As noted earlier in Section 3, the existence of a near minimizer, indeed a global minimizer, is guaranteed by the continuity of $\Gamma_n(\lambda, C_n)$ over the compact set $\lambda(\Upsilon)$. Furthermore, because the conditions for C_n are easily satisfied [for example by the sequence $r_n \log n \to \infty$, where $r_n = o(1)$], it follows that the conditions for Theorem 21 hold in practice, thus establishing the existence of a consistent estimator. Consequently, from our

previous discussion, we can assume with no loss of generality that we have a uniform $o_p(1/\log n)$ estimator at our disposal.

THEOREM 22. *Let $\tilde{\theta}_n$ be a uniform $o_p(1/\log n)$ estimator for θ over a regular sequence of families Υ_n. Let*

$$\tilde{C}_n = \frac{1}{\tilde{\theta}_n}\log(n^{1/(2d+1)})$$

and suppose that the estimator $\hat{\lambda}_n = (\hat{\theta}_n, \hat{\phi}_n) \in \lambda(\Upsilon)$ satisfies

$$\Gamma_n\big(\hat{\lambda}_n, \tilde{C}_n\big) \le O_p(n^{-2(1+d)/(2d+1)}) + \inf_{\lambda\in\lambda(\Upsilon)} \Gamma_n\big(\lambda, \tilde{C}_n\big) \quad \textit{uniformly over } \Upsilon_n.$$

Then $\hat{\theta}_n$ is a uniform $O_p(n^{-d/(2d+1)})$ estimator for θ over Υ_n, where $0 < d \le 1$.

The idea behind the proof of Theorem 22 will be to closely approximate the criterion function $\Gamma_n(\cdot, \tilde{C}_n)$ by a function H_n^2 that will be easier to work with. The approximation will ensure that the near minimizer $\hat{\lambda}_n$ of $\Gamma_n(\cdot, \tilde{C}_n)$ will also nearly minimize H_n^2. We construct H_n^2 so that it achieves its minimum at λ_n; consequently, the distance between $\hat{\lambda}_n$ and λ_n can be determined by the behavior of H_n^2 near its minimum. In essence, H_n^2 plays the role of a metric in determining the rate for $\hat{\lambda}_n$.

The argument behind this heuristic can be formalized through the following lemma.

LEMMA 23 [Compare with Pollard (1993)]. *Suppose that $\{H_n^2\}$ is a sequence of functions and $\{R_n^2\}$ and $\{C_n\}$ are sequences such that the following hold*:

(i) $\hat{\lambda}_n \in \lambda(\Upsilon)$;
(ii) $\Gamma_n(\hat{\lambda}_n, C_n) \le O_p(r_n^2) + \inf_{\lambda\in\lambda(\Upsilon)} \Gamma_n(\lambda, C_n)$;
(iii) $|\Gamma_n(\lambda, C_n) - H_n(\lambda)^2| \le R_n^2 + 2R_nH_n(\lambda)$ *for* $\lambda \in \lambda(\Upsilon)$;
(iv) *There exists a* $\lambda_n \in \lambda(\Upsilon)$ *such that* $H_n(\lambda_n)^2 = 0$.

Then if $R_n^2 \le O_p(r_n^2)$,

$$H_n\big(\hat{\lambda}_n\big)^2 \le O_p\big(r_n^2\big).$$

PROOF. From (iii) evaluated at $\hat{\lambda}_n \in \lambda(\Upsilon)$,

$$H_n\big(\hat{\lambda}_n\big)^2 \le R_n^2 + 2R_nH_n\big(\hat{\lambda}_n\big) + \Gamma_n\big(\hat{\lambda}_n, C_n\big).$$

By (ii) the right-hand side can be increased further to

$$R_n^2 + 2R_nH_n\big(\hat{\lambda}_n\big) + O_p\big(r_n^2\big) + \Gamma_n(\lambda_n, C_n).$$

Invoke (iii) once more, this time evaluated at λ_n. From assumption (iv) infer that $\Gamma_n(\lambda_n, C_n) \le R_n^2$. Consequently,

$$H_n\big(\hat{\lambda}_n\big)^2 \le 2R_n^2 + 2R_nH_n\big(\hat{\lambda}_n\big) + O_p\big(r_n^2\big),$$

which can also be written as

$$H_n(\hat{\lambda}_n)^2 - 2R_n H_n(\hat{\lambda}_n) - R_n^2 \le R_n^2 + O_p(r_n^2) \le O_p(r_n^2).$$

Complete the square on the left-hand side to find that

$$\left(H_n(\hat{\lambda}_n) - R_n\right)^2 \le O_p(r_n^2).$$

It follows that $H_n(\hat{\lambda}_n) \le O_p(r_n)$. □

A natural candidate for H_n^2 presents itself by considering the contribution from the smoothed data [left-hand side of (15)] to $\Gamma_n(\cdot, \tilde{C}_n)$. The following empirical process notation will make this analysis clearer. Write $\hat{\mathbb{P}}_{\gamma_n}$ for the empirical distribution under sampling from $\mathbb{P}_{\gamma_n}$, where $\gamma_n = (\beta_n, \theta_n, G_n)$ is some sequence in Υ. Let $\mu_{\gamma_n} = n^{1/2}(\hat{\mathbb{P}}_{\gamma_n} - \mathbb{P}_{\gamma_n})$ denote the corresponding empirical process.

Let $\lambda_n = (\theta_n, \phi_n)$ denote the value for $\lambda(\gamma_n)$. From expression (16), rewrite the smoothed data on the left-hand side of (15) as

$$\begin{aligned}\mathbb{P}_{\gamma_n} K(u - X) + \left(\hat{\mathbb{P}}_{\gamma_n} - \mathbb{P}_{\gamma_n}\right) K(u - X) \\ = \phi_n\left[K * m_{\theta_n}\right](u) + \Delta_{\gamma_n}(u) + n^{-1/2}\mu_{\gamma_n} K(u - X).\end{aligned}$$

Therefore, substituting this expression, rewrite $\Gamma_n(\lambda, \tilde{C}_n)$ for $\lambda = (\theta, \phi)$ as

$$\begin{aligned}\int_{-\infty}^{-\tilde{C}_n} \Big(\big(\phi_n\left[K * m_{\theta_n}\right] - \phi\left[K * m_\theta\right]\big)(u) \\ + \Delta_{\gamma_n}(u) + n^{-1/2}\mu_{\gamma_n} K(u - X)\Big)^2 du.\end{aligned} \tag{24}$$

The first term inside the square of the integrand is minimized at $\lambda = \lambda_n$. Furthermore, we will see that its contribution to the integral will be larger than the contribution from either of the two remaining terms except when $\lambda \approx \lambda_n$. Thus, an obvious candidate for the H_n^2 in Lemma 23 is the function $H(\cdot, \lambda_n)^2$ defined by

$$H(\lambda, \lambda_n)^2 = \int_{-\infty}^{-\tilde{C}_n} \left(\phi_n\left[K * m_{\theta_n}\right] - \phi\left[K * m_\theta\right]\right)(u)^2 du. \tag{25}$$

Expand the square in (24), subtract H^2, and take absolute values. Now apply the Cauchy–Schwarz inequality to the absolute value of the cross product term to obtain an expression similar to condition (iii) of Lemma 23:

$$\left|\Gamma_n(\lambda, \tilde{C}_n) - H(\lambda, \lambda_n)^2\right| \le R_{\gamma_n}(\tilde{C}_n)^2 + 2R_{\gamma_n}(\tilde{C}_n) H(\lambda, \lambda_n), \tag{26}$$

where $R_{\gamma_n}(\cdot)^2$ is defined by

$$(27)\quad R_{\gamma_n}(C)^2 = \int_{-\infty}^{-C} \left(\Delta_{\gamma_n}(u) + n^{-1/2}\mu_{\gamma_n} K(u - X)\right)^2 du \quad \text{for each } C \geq 0.$$

Now we can use Lemma 23 to prove Theorem 22.

PROOF OF THEOREM 22. Suppose that we are sampling under the model $\mathbb{P}_{\gamma_n}$ defined as above. Take the H_n^2 and R_n^2 of Lemma 23 to be $H(\cdot, \lambda_n)^2$ and $R_{\gamma_n}(\tilde{C}_n)^2$ defined by (25) and (27), respectively. Therefore, by (26), all the conditions of the lemma will be satisfied if we can show that $R_n^2 \leq O_p(r_n^2)$, where $r_n = n^{-(1+d)/(2d+1)}$. The rate will then follow by solving for $\hat{\lambda}_n$ in

$$(28)\quad H_n\left(\hat{\lambda}_n\right)^2 = H\left(\hat{\lambda}_n, \lambda_n\right)^2 \leq O_p\left(r_n^2\right) = O_p(\exp(-2\theta_n C_n))O_p\left(\delta_n^2\right),$$

where $\delta_n = n^{-d/(2d+1)}$ and

$$C_n = \frac{1}{\theta_n}\log\left(n^{1/(2d+1)}\right).$$

The $o_p(1/\log n)$ consistency of $\tilde{\theta}_n$ ensures that $\tilde{C}_n = C_n + o_p(1)$, uniformly over Υ_n. Therefore, with no loss of generality, we can argue along the set $\{\tilde{C}_n \geq C_n - \varepsilon\}$ which occurs with uniform probability tending to 1 over Υ_n, where $\varepsilon > 0$ is some fixed number. Therefore, replace $\tilde{C}_n$ with $C_n - \varepsilon$ to obtain an upper bound to R_n^2. Consequently, by Lemma 30 (see the Appendix),

$$\begin{aligned} R_n^2 &\leq R_{\gamma_n}(C_n - \varepsilon)^2 \\ &= O\left(\exp\left(-2(1+d)\theta_n C_n\right)\right) + O_p\left(n^{-1}\exp\left(-\theta_n C_n\right)\right) \\ &= O\left(n^{-2(1+d)/(2d+1)}\right) + O_p\left(n^{-2(1+d)/(2d+1)}\right) \\ &= O_p\left(r_n^2\right) \quad \text{uniformly over } \Upsilon_n. \end{aligned}$$

To complete the proof we establish the rate by solving for $\hat{\lambda}_n$ in (28). Recall from identity (17) that $[K * m_\theta](u) = \exp(\theta u)(1 - \exp(-\theta))$ when $u \leq 0$. Using this identity and changing variables from u to $u + \tilde{C}_n$, we get

$$\begin{aligned} H_n\left(\hat{\lambda}_n\right)^2 &= H\left(\hat{\lambda}_n, \lambda_n\right)^2 \\ &= B_n^2 \exp\left(-2\theta_n \tilde{C}_n\right)\int_{-\infty}^{0} \exp(2\theta_n u)\left[1 - \rho_n \exp\left(\left(\hat{\theta}_n - \theta_n\right)u\right)\right]^2 du, \end{aligned}$$

where $B_n = \phi_n(1 - \exp(-\theta_n))$ and

$$\rho_n = \frac{\hat{\phi}_n\left(1 - \exp\left(-\hat{\theta}_n\right)\right)}{\phi_n(1 - \exp(-\theta_n))}\exp\left(\left(\theta_n - \hat{\theta}_n\right)\tilde{C}_n\right).$$

The sequence B_n is strictly bounded away from zero by the positivity of $\lambda(\Upsilon)$ (see Lemma 20). Replace $\exp(2\theta_n u)$ by the lower bound $\exp(2\theta^* u)$, where $\theta^* \geq \theta(\Upsilon)$ is a finite upper bound to θ. Therefore, because $\tilde{C}_n = C_n + o_p(1)$, we can invert (28) by solving for $\hat{\lambda}_n$ in

$$\int_{-\infty}^{0} \exp(2\theta^* u)\left[1 - \rho_n \exp\left(\left(\hat{\theta}_n - \theta_n\right)u\right)\right]^2 du \leq O_p\left(\delta_n^2\right). \tag{29}$$

From this we see immediately that $\rho_n = 1 + O_p(\delta_n)$ and that $(\hat{\theta}_n - \theta_n) = O_p(\delta_n)$. Furthermore, because the probability on the right-hand side of (29) holds uniformly over Υ_n, deduce that the rate for $\hat{\theta}_n$ holds uniformly. □

APPENDIX

PROOF OF LEMMA 20. The bounds to θ follow by Assumption 19(i). Therefore, we only need to verify the bounds on ϕ. Suppose that $G \in \mathcal{G}_M((1 + d)+)$. Apply Jensen's inequality to the convex function $F_1(x) = x^{-1/(1+d)}$ over positive x, to show that

$$\begin{aligned} G\exp(-Y) &= GF_1(\exp((1 + d)Y)) \\ &\geq F_1(G\exp((1 + d)Y)) \geq (1 + M)^{-1/(1+d)}. \end{aligned}$$

One more application of Jensen's inequality, this time applied to the convex function $F_2(x) = x^{1+d}$ over positive x, shows that

$$F_2(G\exp(-Y)) \leq GF_2(\exp(-Y)) = G\exp(-(1 + d)Y) \leq (1 + M).$$

Take $(1 + d)$th roots on the left- and right-hand sides. This and the previous inequality show that

$$p_0^* = (1 + M)^{-1/(1+d)} \leq G\exp(-Y) \leq p_1^* = (1 + M)^{1/(1+d)}.$$

The distribution which concentrates at either $\pm\log(1 + M^*)/(1 + d)$ lies in $\mathcal{G}_M((1 + d)+)$ for $0 \leq M^* \leq M$. By varying M^*, deduce that $\{G\exp(-Y)$: $G \in \mathcal{G}_M((1 + d)+)\} = [p_0^*, p_1^*]$.

Now use Assumption 19(ii), to deduce that

$$0 < p_0 = p_0^*\left(\inf_{|\beta| \leq b_0} \mathbb{P}_Z \exp(-\beta' Z)\right)$$

and

$$p_1 = p_1^*\left(\sup_{|\beta| \leq b_0} \mathbb{P}_Z \exp(-\beta' Z)\right) < \infty.$$

This gives the lower and upper bounds to the parameter space. □

LEMMA 30. *Suppose that* $\gamma_n = (\beta_n, \theta_n, G_n)$ *is a sequence in* Υ. *Then for any positive sequence* $C_n \to \infty$,

$$R_{\gamma_n}(C_n)^2 \le O\big(\exp(-2(1+d)\theta_n C_n)\big) + O_p\big(n^{-1}\exp(-\theta_n C_n)\big)$$

uniformly over Υ.

PROOF. The behavior of $R_{\gamma_n}(C_n)^2$ depends upon the sum of the two terms

$$\int_{-\infty}^{-C_n} \Delta_{\gamma_n}(u)^2\,du + \int_{-\infty}^{-C_n} \big(n^{-1/2}\mu_{\gamma_n}K(u-X)\big)^2\,du,$$

which we denote, from left to right, as (A) and (B). We consider the contribution from each of these terms separately.

(A) Use the bound (14) and change variables from x to $u - x$ in $\Delta_{\gamma_n}(u)$ to see that when $u \le 0$,

$$\begin{aligned}\Delta_{\gamma_n}(u) &= \int\int_0^1 f_2(u-x, z|\gamma_n)\,dx\,d\nu(z)\\ &\le (1+M)\exp\big((1+d)\theta_n u\big)\\ &\quad\times \int_0^1 \theta_n \exp\big(-(1+d)\theta_n x\big)\,dx \int h(z)\exp\big(-(1+d)\beta_n' z\big)\,d\nu(z).\end{aligned}$$

The first integral in the last inequality is bounded by 1, while the second integral is uniformly bounded by Assumption 19(ii). Deduce that

$$\int_{-\infty}^{-C_n} \Delta_{\gamma_n}(u)^2\,du \le O\big(\exp(-2(1+d)\theta_n C_n)\big) \quad \text{uniformly over } \Upsilon.$$

(B) Expand the integrand and take its expectation under $\mathbb{P}_{\gamma_n}^n$. The cross-product terms factorize by independence and vanish because they have zero means. Bound the remaining term by

$$\frac{1}{n}\int_{-\infty}^{-C_n} \mathbb{P}_{\gamma_n}K(u-X)^2\,du = \frac{1}{n}\int_{-\infty}^{-C_n} \phi_n\big[K * m_{\theta_n}\big](u) + \frac{1}{n}\int_{-\infty}^{-C_n} \Delta_{\gamma_n}(u). \tag{31}$$

[The right-hand side follows from the identity (16), the fact that $K^2 = K$ for a uniform(0, 1) kernel and by setting $\lambda(\gamma_n) = (\theta_n, \phi_n)$.]

We have already evaluated the behavior of Δ_{γ_n} in the second integral on the right-hand side of (31). The integrand of the first integral is uniformly $O(\exp(\theta_n u))$ by identity (17) and the boundedness of $\lambda(\Upsilon)$. Deduce by Chebyshev's inequality that (B) is $O_p(n^{-1}\exp(-\theta_n C_n))$, uniformly over Υ. □

REFERENCES

BEGUN, J. M., HALL, W. J., HUANG, W-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

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. (1995). Uniform rates of estimation in the semiparametric Weibull mixture model. Technical Report 278, Laboratory of Research in Statistics and Probability, Carleton Univ.–Univ. Ottawa.

ISHWARAN, H. (1996). Identifiability and rates of estimation for scale parameters in location mixture models. *Ann. Statist.* **24** 1560–1571.

POLLARD, D. (1993). The asymptotics of a binary choice model. *Econometrica*. To appear.

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