The Annals of Statistics
1999, Vol. 27, No. 1, 159–177

INFORMATION IN SEMIPARAMETRIC MIXTURES OF EXPONENTIAL FAMILIES[1]

By Hemant Ishwaran

University of Ottawa

In a class of semiparametric mixture models, the score function (and consequently the effective information) for a finite-dimensional parameter can be made arbitrarily small depending upon the direction taken in the parameter space. This result holds for a broad range of semiparametric mixtures over exponential families and includes examples such as the gamma semiparametric mixture, the normal mean mixture, the Weibull semiparametric mixture and the negative binomial mixture. The near-zero information rules out the usual parametric $\sqrt{n}$ rate for the finite-dimensional parameter, but even more surprising is that the rate continues to be unattainable even when the mixing distribution is constrained to be countably discrete. Two key conditions which lead to a loss of information are the smoothness of the underlying density and whether a sufficient statistic is invertible.

1. Introduction. This paper studies the loss of information associated with estimating a finite-dimensional parameter θ due to the presence of an infinite-dimensional nuisance mixing distribution G. The models that will be considered here are finite semiparametric mixtures over exponential families. One of the key results of the paper shows a loss of information for θ in models which satisfy a smoothness condition and which contain an invertible sufficient statistic. Semiparametric mixture models are typically nonidentified without some form of constraint, so a loss of information by itself is not too surprising. However, the loss of information seen here persists even for models which are constrained to allow for only discrete mixtures. We will see that allowing discrete mixtures to have limit points will lead to a breakdown of the classical $\sqrt{n}$ inference for θ.

The exact details of the problem are as follows. For real-valued $(\theta, y) \in \Theta \otimes \mathscr{Y}$, let

$$f(x|\theta, y) = \exp(ys(x,\theta) + t(x,\theta) + b(\theta, y)) \tag{1}$$

denote a density taken with respect to a σ-finite measure λ on a measurable space $(\mathscr{X}, \mathscr{B})$, where for each $\theta \in \Theta$, the maps $s(\cdot, \theta)$ and $t(\cdot, \theta)$ are Borel measurable functions from $(\mathscr{X}, \mathscr{B})$ onto $\mathbb{R}$. For fixed θ, densities of the form (1) define a one-parameter exponential family indexed by y, with natural

Received August 1996; revised August 1998.

[1] Supported in part by grant funds from the Natural Sciences and Engineering Research Council of Canada.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62G20, 62B10.

Key words and phrases. Semiparametric mixture, mixture model, structural parameter, information.

parameter space

$$\mathscr{Y}(\theta) = \left\{ y\colon 0 < \int \exp(ys(x,\theta) + t(x,\theta))\, d\lambda(x) < \infty \right\}.$$

It will be assumed that $\mathscr{Y}(\theta) = \mathscr{Y}$ for each θ and that $\mathscr{Y}$ is nonempty. By the convexity of the natural parameter space, this guarantees that $\mathscr{Y}$ is either a nonempty interval in $\mathbb{R}$ or all of $\mathbb{R}$.

Let G be some unspecified mixing distribution on $\mathscr{Y}$ and $f(x|\theta, y)$ a density of the form (1), where $f(x|\theta, y)$ is assumed to be measurable in (x, y). Then

$$\begin{aligned} f(x|\theta,G) &= \int f(x|\theta,y)\, dG(y) \\ &= \int \exp(ys(x,\theta) + t(x,\theta) + b(\theta,y))\, dG(y) \end{aligned} \tag{2}$$

is a semiparametric mixture over a density from an exponential family.

The problem for estimating θ in semiparametric mixtures of the form (2) has been studied by Lindsay (1983) and Van der Vaart (1988). Both authors have considered examples in which the measure of information for θ remains positive in the presence of the mixing distribution G. Van der Vaart (1988) constructed estimators for θ which are efficient under the assumption that a least-favorable submodel exists. The existence of such submodels are shown to hold in models like (2) for discrete G with limit points. The proof relies on the completeness of exponential families.

The completeness property of the exponential family will also create a loss of information for θ. By modifying some of the arguments given in Van der Vaart (1988), we will be able to characterize models in which such a loss of information occurs. In particular, we will present conditions on G and the underlying densities (1) which cause the information for θ to be arbitrarily small. As in Begun, Hall, Huang and Wellner (1983), and Lindsay (1983), we measure the information for θ by "the effective score for θ," which reflects the reduction in the Fisher information for θ due to the presence of the nuisance parameter G. In particular, we will study mixtures of the form (2) where the effective information for θ approaches zero even when G is constrained to be discrete.

The problem can be further motivated by the following example.

EXAMPLE (Normal mean mixture). If Z is a standard normal variable independent of $Y \sim G$, then $X = \sqrt{\theta}Z + Y$ is a normal mean mixture with unknown variance $\theta > 0$. This class of models is of the form (2) because the density for $(X|\theta, y) \sim N(y, \theta)$ can be written as in (1),

$$f(x|\theta,y) = \exp\left(y\frac{x}{\theta} - \frac{x^2}{2\theta} + \left[-\frac{1}{2}\log(2\pi\theta) - \frac{y^2}{2\theta} \right] \right),$$

where $\Theta \otimes \mathscr{Y} = \mathbb{R}^+ \otimes \mathbb{R}$.

It is well known that the normal mean mixture is identified only when G contains no normal components [Kiefer and Wolfowitz (1956)]. Without this constraint, the nonidentifiability of the model is easily demonstrated by noting that

$$(X = Z) \overset{\mathscr{D}}{=} (X = \sqrt{1-\theta}Z + Y),$$

where $Y \sim N(0, \theta)$ is independent of Z.

Clearly then, there is no information for θ in the unconstrained model or at least not for inferential problems that are one-sided such as in testing a null hypothesis like $\theta \le \theta_0$. Thus, without constraints we cannot estimate θ at any rate, but what if we apply constraints to the model? Will there still be a loss of information, and will the $\sqrt{n}$ rate remain unattainable even with a strongly identified model? The answer to this question is "yes," because we will find that even when G is constrained to be discrete there is still a loss of information for θ that precludes a $\sqrt{n}$ rate of estimation. Furthermore, this phenomenon persists even when $\theta \ge \theta_0$, which shows that the loss of information in the discrete case is more than just an approximation to what occurs in the continuous problem.

One might conjecture that the $\sqrt{n}$ rate for θ is always unachievable in finite semiparametric mixtures like (2). Such a conjecture seems reasonable given the parallel result, which is that the $\sqrt{n}$ rate fails to hold when estimating G in finite mixtures [Chen (1995)]. However, the classical breakdown does not always occur for θ. Consider semiparametric mixtures $N(\theta, Y)$, where $Y > 0$ is a scale random variable and $\theta \in \mathbb{R}$ is the unknown mean. Then

$$f(x|\theta, y) = \exp\left(-\frac{1}{2y}(x-\theta)^2 - \frac{1}{2}\log(2\pi y)\right),$$

which can be rewritten as (1) by reparameterizing y as y^{-1}. However, the classical rate fails to break down in this example because θ can always be estimated at a $\sqrt{n}$ rate using the median of the data.

What sets this example apart from the normal mean mixture model is the form for the sufficient statistic $s(\cdot, \theta)$ for y. As we will see, a critical assumption leading to a loss of information will depend upon $s(\cdot, \theta)$ being invertible. In the normal mean mixture model $s(x, \theta) = x/\theta$ is 1:1, while in the previous example $s(x, \theta) = -(x-\theta)^2/2$, which is not uniquely invertible over $\mathbb{R}$.

1.1. *Further examples.* Our problem can be further motivated by some additional examples. The next example is interesting because the loss of information for θ appears to exist in only one direction. This one-sidedness is also the reason why our regularity conditions (condition C1 in Section 2) are unidirectional.

EXAMPLE (Weibull semiparametric mixture). Let W denote a standard exponential variable, independent of a positive variable $Y \sim G$. Then the Weibull semiparametric mixture is the distribution for the variable $X = (W/Y)^{1/\theta}$. Here $\theta > 0$ acts as a shape parameter and $Y > 0$ as an unobserved scale mixing variable. If θ and $Y = y$ were known, X would have the Weibull(θ, y) distribution with density

$$f(x|\theta, y) = \exp\big(y(-x^{\theta}) + (\theta - 1)\log x + \log(\theta y)\big), \qquad x > 0.$$

This has the same structure as (1) with $\Theta \otimes \mathscr{Y} = \mathbb{R}^{+} \otimes \mathbb{R}^{+}$.

Jewell (1982) observed that a Weibull(θ', 1) can always be reexpressed as a mixture of Weibull(θ, y) variables, if $\theta > \theta'$. That is, there exists a Y independent of W such that

$$\text{Weibull}(\theta', 1) =_{\mathscr{D}} \text{Weibull}(\theta, Y), \qquad \theta > \theta'.$$

An explicit construction for Y is given in Ishwaran (1994), who shows how moment constraints for G translate into slower than $\sqrt{n}$ rates for θ. Although the moment constraints on G are enough to ensure that the models are identified, they still do not guarantee the $\sqrt{n}$ rate. Here we will show that the problem persists even for discrete G.

Another example in which there seems to be a one-sided lack of information (at least for the continuous case) is the gamma semiparametric mixture.

EXAMPLE (Gamma semiparametric mixture). With shape parameter $\theta > 0$ and scale parameter $y > 0$, the Gamma(θ, y) has density

$$\begin{aligned} f(x|\theta, y) &= \Gamma(\theta)^{-1} y^{\theta} x^{\theta - 1} \exp(-yx) \\ &= \exp\big(y(-x) + (\theta - 1)\log x + [\theta \log y - \log \Gamma(\theta)]\big), \qquad x > 0, \end{aligned}$$

where Γ is the gamma function. Clearly, the density is of the form (1). Here $\Theta \otimes \mathscr{Y} = \mathbb{R}^{+} \otimes \mathbb{R}^{+}$.

Unconstrained, the model is unidentified and hence contains no information for θ. The lack of identification follows from an exercise in Lindsay [(1995), page 55],

$$f(x|\theta', y') = \int f(x|\theta, y)\, dG(y) \quad \text{for } \theta > \theta',$$

where $dG(y) \propto (y - y')^{\theta - \theta' - 1} y^{-\theta}\{y \geq y'\}$.

As we will see, the $\sqrt{n}$ rate will fail to hold in the normal, Weibull and gamma mixtures described above. Because many other mixtures can be described in terms of these models, the parametric rate must fail to hold in a wider range of models than has been commonly recognized or appreciated. For example, the same failure in the gamma mixture must also exist in the inverse-gamma mixture and the scaled inverse-chi-square mixture. Another important example in which the $\sqrt{n}$ rate fails is the negative binomial mixture. Its distribution arises through mixing with the gamma distribution.

EXAMPLE (Negative binomial mixture). The negative binomial, Neg-Bin(θ, β), is the distribution for a Poisson variable whose rate parameter has the Gamma(θ, β) distribution. The additional parameter makes it a robust alternative to the Poisson distribution. When $\theta > 0$ and $\beta > 0$ are known, the Neg-Bin(θ, β) has density

$$f(x|\theta,\beta) = \frac{\Gamma(x+\theta)}{\Gamma(x+1)\Gamma(\theta)}\left(\frac{\beta}{1+\beta}\right)^{\theta}\left(\frac{1}{1+\beta}\right)^{x} \quad \text{for } x = 0, 1, \dots .$$

Under the 1:1 transformation $y = -\log(1+\beta)$, the density can be reexpressed as (1) with $s(x,\theta) = x$ and

$$t(x,\theta) = \log\left(\frac{\Gamma(x+\theta)}{\Gamma(x+1)\Gamma(\theta)}\right),$$
$$b(\theta, y) = \theta[y + \log(\exp(-y) - 1)].$$

[Note that under the alternate parameterization, $y = \beta/(\beta+1)$, the distribution represents the number of failures before the θth success, where the probability of success is y.]

We will use the parameterization $y = -\log(1+\beta)$ in studying mixtures of the Neg-Bin(θ, y). In this case, $\Theta \otimes \mathscr{Y} = \mathbb{R}^{+} \otimes \mathbb{R}^{-}$ and λ is counting measure on $\mathscr{X} = \{0, 1, \dots\}$ over the σ-algebra $2^{\mathscr{X}}$, the set of all subsets of $\mathscr{X}$.

The loss of information that we will study will be intimately tied to the discreteness assumed for G. To make this more precise we need to introduce some notation. For each integer $j = 1, 2, \dots,$ let $\mathscr{G}_j$ denote the class of discrete distributions that assign positive mass to exactly j atoms on $\mathscr{Y}$. Set $\mathscr{G}_F = \bigcup_{j=1}^{\infty} \mathscr{G}_j$ to be the class of all (finitely) discrete mixtures, and define $\mathscr{P}(\Theta, \mathscr{G}_F)$ to be the class of mixtures of the form (2) over $\theta \in \Theta$ and $G \in \mathscr{G}_F$.

This paper will show that the $\sqrt{n}$ rate for θ fails to hold if the class of finite mixing distributions $\mathscr{G}_F$ is slightly expanded to include all discrete distributions. Let $\mathscr{G}_\infty$ be the class of discrete distributions that have a countable support on $\mathscr{Y}$ (the support for a distribution is the smallest set with measure one). Let $\mathscr{G}_D = \mathscr{G}_F \cup \mathscr{G}_\infty$. Our main result, Theorem 3, describes conditions on the underlying densities (1) which ensure that the θ parameter in the class of models $\mathscr{P}(\Theta, \mathscr{G}_D)$ is not estimable at a $\sqrt{n}$ rate. Indeed, we will find that the rate remains unattainable even if the G are constrained to share the same support, as long as it is a countable set containing a limit point. These results will hold for the motivating examples presented here (Section 3 verifies the conditions of Theorem 3 for the four examples).

Although we find a loss of information in each of our examples, the models under study are all strongly identified, at least for the case when $G \in \mathscr{G}_F$. This result is given in Section 4, which presents conditions sufficient for identification in finite semiparametric mixtures of the form (2). The theorem is applied to five different models, including the four described above.

2. Zero information.

Let $\mathscr{P}(\Theta, \mathscr{G}_D)$ be the class of mixtures with densities of the form (2), where $\mathscr{G}_D$ is the class of discrete distributions on $\mathscr{Y}$. We will exhibit fairly simple conditions which lead to a loss of information and which preclude a $\sqrt{n}$ rate of estimation for the θ parameter. The conditions will depend upon the smoothness of the underlying densities (1) and whether the sufficient statistic $s(\cdot, \theta)$ for y is invertible.

Suppose $P_0 \in \mathscr{P}(\Theta, \mathscr{G}_D)$ is a mixture with structural parameter $\theta_0 \in \Theta$ and mixing distribution $G_0 \in \mathscr{G}_D$. Write f_0 for its density taken with respect to λ. The failure of the $\sqrt{n}$ rate will be based upon a construction which will depend upon the structure of G_0. Suppose that the support for G_0 is countably discrete with a limit point residing in the interior of $\mathscr{Y}$. Then we will show that for each $\varepsilon > 0$, there exists a $G_\tau \in \mathscr{G}_D$, for small τ, such that

$$(3) \qquad P_0\left(\frac{f_\tau}{f_0} - 1\right)^2 \approx \varepsilon\tau^2,$$

where

$$(4) \qquad f_\tau(x) = \int f(x|\theta_0 + \tau, y)\, dG_\tau(y).$$

In particular, the construction describes a one-dimensional approach $(\theta_0 + \tau, G_\tau)$ to (θ_0, G_0) through the parameter space $\Theta \otimes \mathscr{G}_D$ which causes the score function (and hence the information) for θ to become arbitrarily small. The relationship (3) will then establish the rate assertion by a standard Hellinger distance argument.

The G_τ used in the construction is defined by

$$(5) \quad dG_\tau = dG_0(1 + \tau h_\tau) \quad \text{where } G_0|h_\tau| = \int |h_\tau(y)|\, dG_0(y) = o(1/\tau).$$

That is, $1 + \tau h_\tau$ is the Radon–Nikodym derivative of G_τ with respect to G_0. The construction will yield a fairly strong result, because it relies on a G_τ which is absolutely continuous with respect to G_0, and which converges to G_0 in the total variation distance. Slightly rephrased, because G_τ and G_0 share the same support, this means that even if we knew exactly what the support was, we still would not be able to estimate θ at the usual parametric $\sqrt{n}$ rate.

The heuristic for the construction is as follows. Write $L_0^2(G_0)$ for the equivalence class of G_0-square integrable functions with zero G_0 expectation. Take some $h \in L_0^2(G_0)$, truncate the function and then center it by its G_0 expectation. This gives an h_τ that can be used in constructing a G_τ as in (5). Now assuming that our densities are smooth enough to be differentiated, we can expand (4) so that

$$\begin{aligned} f_\tau(x) &= \int f(x|\theta_0 + \tau, y)[1 + \tau h_\tau(y)]\, dG_0(y) \\ &\approx \int \left[f(x|\theta_0, y) + \tau \frac{\partial}{\partial\theta} f(x|\theta_0, y) + \cdots \right] [1 + \tau h(y)]\, dG_0(y) \\ &\approx f_0(x) + \tau \int \left[h(y) f(x|\theta_0, y) + \frac{\partial}{\partial\theta} f(x|\theta_0, y) \right] dG_0(y). \end{aligned}$$

Then, dividing throughout by f_0,

$$\frac{f_\tau(x)}{f_0(x)} \approx 1 + \tau[A(G_0, h)(x) + \rho(x)], \tag{6}$$

where

$$A(G_0, h)(x) = \frac{1}{f_0(x)} \int h(y) f(x|\theta_0, y)\, dG_0(y) \tag{7}$$

and

$$\rho(x) = \frac{1}{f_0(x)} \int \frac{\partial}{\partial\theta} f(x|\theta_0, y)\, dG_0(y). \tag{8}$$

To establish the slower than $\sqrt{n}$ rate, the aim will be to find a $G_0 \in \mathscr{G}_D$ and an $h \in L_0^2(G_0)$ such that

$$A(G_0, h)(x) \approx -\rho(x).$$

This will then show by (6) that $f_\tau \approx f_0$, which will complete the argument using (3).

Here are the regularity conditions that formalize the heuristic.

Condition C1. For small τ, either positive or negative, but not necessarily both:

(i) The derivative

$$\Delta(x, y, \theta_0 + \tau) = \frac{\partial}{\partial\theta} f(x|\theta_0 + \tau, y)$$

exists for a.a. $(x, y)[\lambda \otimes G_0]$.

(ii) $\Delta(x, y, \theta_0 + \tau)$ is continuous in τ.

(iii) There exists a dominating function M such that

$$\frac{\Delta(x, y, \theta_0 + \tau)^2}{f(x|\theta_0, y)} \le M(x, y) \quad \text{where } M \in L^1(\lambda \otimes G_0). \tag{9}$$

With condition C1, we can formalize the steps in the heuristic in proving the following.

Lemma 1. *Assume that condition* C1 *holds for* $P_0 \in \mathscr{P}(\Theta, \mathscr{G}_D)$, *where* P_0 *has parameter* (θ_0, G_0). *Then for each* $h \in L_0^2(G_0)$, *there exists a* $G_\tau \in \mathscr{G}_D$ *of the form* (5) *such that*

$$P_0\left(\frac{f_\tau}{f_0} - 1\right)^2 \le \tau^2 P_0(A(G_0, h) + \rho)^2 + o(\tau^2), \tag{10}$$

where $\rho \in L_0^2(P_0)$ *is defined by* (8) *and* $A(G_0, \cdot\,)$ *is the linear map from* $L_0^2(G_0)$ *into* $L_0^2(P_0)$ *defined by* (7).

The next lemma describes the condition needed for G_0 in order to make the first term on the right-hand side of (10) arbitrarily small. The proof of the lemma relies on the completeness of exponential families and requires $s(\cdot, \theta)$ to be invertible.

LEMMA 2. *Suppose the support for $G_0 \in \mathscr{G}_D$ contains a limit point which lies in the interior of $\mathscr{Y}$. If $s(\cdot, \theta_0)$ is bimeasurable* (1:1 *on $\mathscr{X}$ with a measurable inverse*), *then $\{A(G_0, h): h \in L_0^2(G_0)\}$ is dense in $L_0^2(P_0)$.*

Lemmas 1 and 2 yield the slower than $\sqrt{n}$ rate (their proofs can be found in the Appendix).

THEOREM 3. *Assume that condition* C1 *holds for $P_0 \in \mathscr{P}(\Theta, \mathscr{G}_D)$ with parameter (θ_0, G_0), where G_0 and $s(\cdot, \theta_0)$ satisfy the conditions of Lemma* 2. *Let $\mathscr{G}_0 \subseteq \mathscr{G}_D$ be the class of discrete distributions with the same support as G_0. If $X_1, X_2, \dots$ is an independent sequence from P_0, then an estimator $\hat{\theta}_n(X_1, X_2, \dots, X_n)$ for θ_0 must have rate of convergence slower than $O_p(n^{-1/2})$ in the class of mixtures $\mathscr{P}(\Theta, \mathscr{G}_0)$.*

PROOF. The operator $A(G_0, \cdot\,)$ is dense by Lemma 2. Therefore, for each $\varepsilon > 0$ there exists a function $h \in L_0^2(G_0)$ satisfying $P_0(A(G_0, h) + \rho)^2 \le \varepsilon$. Hence, by Lemma 1 there exists a $G_\tau \in \mathscr{G}_0$ so that

$$\int \frac{(f_\tau - f_0)^2}{f_0} \le \varepsilon\tau^2 + o(\tau^2) \quad \text{eventually.}$$

This bounds the Hellinger distance because $(\sqrt{a} - \sqrt{b}\,)^2 \le (a-b)^2/b$ for any $a, b \ge 0$. Therefore, we have exhibited mixtures whose θ values are separated by τ, but whose Hellinger distance is bounded by $\varepsilon\tau^2$ for arbitrarily small $\varepsilon > 0$. Furthermore, these mixtures lie in $\mathscr{P}(\Theta, \mathscr{G}_0)$. The theorem now follows from standard results concerning the Hellinger distance and its role in determining rates of estimation [see Le Cam (1973)]. □

3. Examples of slower than $\sqrt{\mathbf{n}}$ estimation.

Here we establish the non-$\sqrt{n}$ rate of estimation for θ in each of our four motivating examples. We show this by checking the conditions in Theorem 3, which, as we will see, are quite straightforward, excepting condition (9). Interestingly, in at least one of our examples (the Weibull) the dominating condition (9) appears to hold only for values of θ in a particular direction: that is, for τ values that are either strictly positive or strictly negative. Because the Weibull model is known to have a one-sided loss of information, it would appear that condition (9) can act like a diagnostic for identifying such models.

EXAMPLE (Normal mean mixture). The continuity of the derivative is straightforward. Furthermore, because $s(x, \theta) = x/\theta$ is 1:1 with a Borel measurable inverse, we need only check condition (9) in order to establish the non-$\sqrt{n}$ rate.

Simple differentiation shows that

$$\Delta(x, y, \theta) = \left[-\frac{1}{2\theta} + \frac{1}{2\theta^2}(x-y)^2 \right] f(x|\theta, y).$$

With $\theta_0 + \varepsilon \geq \theta \geq \theta_0$, we obtain the bound

$$\frac{\Delta(x, y, \theta)^2}{f(x|\theta_0, y)} \leq \left[\frac{1}{2\theta_0} + \frac{1}{2\theta_0^2}(x-y)^2 \right]^2 \frac{1}{\sqrt{2\pi\theta_0}} \exp\left(-\frac{(\theta_0 - \varepsilon)(x-y)^2}{2(\theta_0 + \varepsilon)\theta_0} \right).$$

If $M(x, y)$ is the function on the right-hand side of the bound, then it easily follows that $[G_0 \otimes \lambda]M < \infty$ for a small enough $\varepsilon > 0$. This verifies the conditions of the theorem.

It is interesting to note here that the conditions could have also been verified for $\tau < 0$. However, by verifying them for $\tau > 0$ we have shown that there is a loss of information associated with $\theta > \theta_0$. This is interesting because the loss of information in the unconstrained model occurs for $\theta < \theta_0$ when Y has a distribution with a normal component. The loss of information in the other direction signifies a loss of efficiency in the discrete mixture problem, which is unrelated to having G_τ approximating a normal distribution.

EXAMPLE (Negative binomial mixture). Verifying condition (9) is the only tricky thing here [notice that $s(x, \theta) = x$ is invertible with a bimeasurable inverse from $(\mathscr{X}, 2^{\mathscr{X}})$ onto $(\mathscr{X}, 2^{\mathscr{X}})$].

Differentiating,

$$\Delta(x, y, \theta) = \left[\frac{\partial}{\partial\theta} \log\left(\frac{\Gamma(x+\theta)}{\Gamma(\theta)} \right) + y + \log(\exp(-y) - 1) \right] f(x|\theta, y). \tag{11}$$

Observe that the derivative exists and is properly defined due to the continuity of Γ and from the fact that $\exp(-y) - 1 > 0$ over $\mathscr{Y} = \mathbb{R}^-$. Furthermore, the gamma function is analytic, and can be expressed as the infinite product expansion [Ahlfors (1979), Chapter 5.2.4]

$$\Gamma(z) = \frac{1}{z}\exp(-\gamma z) \prod_{j \geq 1} \exp\left(\frac{z}{j} \right) \left(1 + \frac{z}{j} \right)^{-1} \quad \text{for } \mathscr{R}(z) > 0,$$

where γ is Euler's constant with approximate value 0.57722. The product expansion provides a convenient method for bounding (11). Taking the log, and differentiating term by term, we get the following useful expression:

$$\frac{\partial}{\partial z} \log \Gamma(z) = -\gamma - \frac{1}{z} + \sum_{j=1}^{\infty} \frac{z}{j^2} \left(1 + \frac{z}{j} \right)^{-1} \quad \text{for } \mathscr{R}(z) > 0.$$

Therefore, for $x \in \mathscr{X}$ and $\theta > 0$,

$$\begin{aligned} &\frac{\partial}{\partial\theta}\log\left(\frac{\Gamma(x+\theta)}{\Gamma(\theta)}\right) \\ (12)\qquad &\quad = \frac{x}{\theta(x+\theta)} + \sum_{j=1}^{\infty}\frac{x}{j^2}\left(1+\frac{x+\theta}{j}\right)^{-1}\left(1+\frac{\theta}{j}\right)^{-1}, \end{aligned}$$

which is bounded by $1/\theta + x\zeta(2)$, where $\zeta(2) = \sum_{j=1}^{\infty}1/j^2$ is the Riemann zeta function evaluated at 2. With $\theta_0 \geq \theta \geq \theta_0 - \varepsilon$, this gives the following upper bound to the square bracketed term in (11):

$$R(x,y) = \frac{1}{\theta_0 - \varepsilon} + x\zeta(2) + |y + \log(\exp(-y) - 1)|.$$

Because $y + \log(\exp(-y) - 1) < 0$ over $\mathscr{Y}$, we get with $\theta \geq \theta_0 - \varepsilon$,

$$\begin{aligned} \frac{f(x|\theta,y)^2}{f(x|\theta_0,y)} \leq \exp\big(yx + (\theta_0 - 2\varepsilon)[y + \log(\exp(-y) - 1)] \\ + 2t(x,\theta) - t(x,\theta_0)\big). \end{aligned}$$

We can use (12) to expand $t(x,\cdot)$ around θ_0. The mean value theorem gives

$$t(x,\theta) = t(x,\theta_0) + (\theta - \theta_0)\frac{\partial}{\partial\theta}\log\left(\frac{\Gamma(x+\theta^*)}{\Gamma(\theta^*)}\right) \quad \text{where } \theta \leq \theta^* \leq \theta_0.$$

When $\theta < \theta_0$, the second term on the right is negative by the positivity of (12). Therefore, with $\theta_0 \geq \theta \geq \theta_0 - \varepsilon$,

$$\begin{aligned} (13)\qquad &\frac{\Delta(x,y,\theta)^2}{f(x|\theta_0,y)} \\ &\quad \leq R(x,y)^2\exp(-2\varepsilon[y + \log(\exp(-y) - 1)])f(x|\theta_0,y). \end{aligned}$$

There are several things to notice in showing that the right-hand side of (13) is $\lambda \otimes G_0$ integrable. First, the second moment of a Neg-Bin(θ, y) equals

$$\frac{\theta(\exp(-y) + \theta)}{(\exp(-y) - 1)^2},$$

which remains bounded for large $y < 0$, but is $O(y^{-2})$ for small $y < 0$. Secondly, $y + \log(\exp(-y) - 1)$ is bounded for large $y < 0$ and $O(\log(-y))$ for small $y < 0$. Therefore, if $G_0 Y^{-2} < \infty$, then (13) is $\lambda \otimes G_0$ integrable. Hence, the right-hand side of (13) gives the required M for a G_0 with a finite inverse-second moment.

EXAMPLE (Gamma semiparametric mixture). All the conditions of the theorem are straightforward except (9). Simple differentiation yields

$$\begin{aligned} &\frac{\Delta(x,y,\theta)^2}{f(x|\theta_0,y)} \\ &\quad = \frac{\Gamma(\theta_0)}{\Gamma(\theta)^2}\left[\log x + \log y - \frac{\partial}{\partial\theta}\log\Gamma(\theta)\right]^2 (yx)^{2\theta-\theta_0}x^{-1}\exp(-yx). \end{aligned} \tag{14}$$

By considering whether $yx \le 1$ or $yx > 1$, we have for $\theta_0 + \varepsilon/2 \ge \theta \ge \theta_0$,

$$(yx)^{2\theta-\theta_0} \le (yx)^{\theta_0} + (yx)^{\theta_0+\varepsilon}.$$

The terms involving the gamma function in (14) can be bounded so that they do not depend upon θ. This and the previous bound gives us our candidate M function. Its $G_0 \otimes \lambda$ integrability follows by noting that

$$\begin{aligned} &\int_0^\infty (\log x + \log y + C)^2 (yx)^\alpha x^{-1}\exp(-yx)\,dx \\ &\qquad = \int_0^\infty (\log x + C)^2 x^{\alpha-1}\exp(-x)\,dx, \end{aligned}$$

which is finite for $\alpha > 0$.

EXAMPLE (Weibull semiparametric mixture). For the Weibull mixture, note that $s(x,\theta) = -x^\theta$ is 1:1 with a Borel measurable inverse from $\mathbb{R}^-$ onto $\mathbb{R}^+$. To verify condition (9), differentiate to obtain

$$\Delta(x,y,\theta) = \left[(1 - yx^\theta)\log x + \frac{1}{\theta}\right] f(x|\theta,y).$$

With $\theta_0 + \varepsilon \ge \theta \ge \theta_0$,

$$\frac{\Delta(x,y,\theta)^2}{f(x|\theta_0,y)} \le R(x,y,\theta)\,yx^{2\theta-\theta_0-1}\exp(-y(2x^\theta - x^{\theta_0})),$$

where $R(x,y,\theta) = \theta_0^{-1}(\theta_0+\varepsilon)^2[(1+yx^\theta)|\log x| + 1/\theta_0]^2$. The ratio can be further bounded:

$$\begin{aligned} &\frac{\Delta(x,y,\theta)^2}{f(x|\theta_0,y)} \\ &\quad \le \begin{cases} R(x,y,0)\,yx^{\theta_0-1}\exp(y), & \text{if } 0 \le x \le 1, \\ R(x,y,\theta_0+\varepsilon)\,yx^{\theta_0+\varepsilon-1}\exp(-yx^{\theta_0}), & \text{if } x > 1. \end{cases} \end{aligned} \tag{15}$$

Some simple calculus shows that both inequalities are integrable under the constraint that $G_0\exp((1+r)Y) < \infty$, for some $r > 0$. Therefore, by piecing together the inequalities in (15) we can construct an M satisfying (9) when G_0 satisfies the necessary moment constraint.

The calculations given above indicate that condition (9) only holds when $\theta > \theta_0$. This seems to be a signal that in the Weibull model there is loss of information only when $\theta > \theta_0$ and not necessarily when $\theta < \theta_0$. Jewell (1982) and Ishwaran (1994) observed the same one-sided lack of identification.

4. Identification. The identification for the class of finite mixtures $\mathscr{P}(\Theta, \mathscr{G}_F)$ will depend upon how well we can distinguish between densities $f(\cdot|\theta, y)$ and $f(\cdot|\theta', y')$ of the form given in (1). Because of the special structures of (1), the identification will depend upon the behavior of

$$D(x, \theta, \theta', y, y') = ys(x, \theta) + t(x, \theta) - (y's(x, \theta') + t(x, \theta'))$$

over $x \in \mathscr{X}$, $\theta, \theta' \in \Theta$ and $y, y' \in \mathscr{Y}$.

Theorem 4 (given below) presents conditions based on the behavior of $D(x,\theta, \theta', y, y')$ which ensure identification in each of our four motivating examples, as well as an additional example, the inverse Gaussian mixture. It should be noted that Teicher (1963) proved the identification for the normal and gamma finite mixtures [Kiefer and Wolfowitz (1956) also discussed the identification for the normal mean mixture], while Elbers and Ridder (1982), as well as Heckman and Singer (1984), proved identification for the Weibull semiparametric mixture. Although the identification in each of these three models has been addressed, there still does not seem to be a theorem which addresses the identification for all of the five models presented here (to my knowledge this is the first identification proof for the inverse Gaussian and negative binomial mixture). Theorem 4 presents a unified approach for studying all these models. Furthermore, the conditions for the theorem depend only upon the form of the density and are much easier to work with than conditions formulated in terms of the characteristic function [as in Teicher (1963)]. This is especially helpful when the characteristic function is complicated, as in the Weibull and inverse Gaussian case.

THEOREM 4. *For the class of densities of the form* (1), *suppose there exists values* $x_0, x_0^* \in \bar{\mathscr{X}}$ *such that*:

(a) *For one of the cases* $y < y'$ *or* $y > y'$,

$$D(x, \theta, \theta', y, y') \to \begin{cases} +\infty, & \text{if } \theta = \theta' \text{ as } x \to \{x_0, x_0^*\}, \\ \pm\infty, & \text{if } \theta \neq \theta' \text{ as } x \to x_0. \end{cases} \tag{16}$$

(b) *For* $y = y'$ *and one of the cases* $\theta > \theta'$ *or* $\theta < \theta'$,

$$D(x, \theta, \theta', y, y') \to -\infty \quad \text{as } x \to x_0^*. \tag{17}$$

Then $\mathscr{P}(\Theta, \mathscr{G}_F)$ *is identified.*

PROOF. Suppose there exist two mixtures in $\mathscr{P}(\Theta, \mathscr{G}_F)$ so that

$$\sum_{i=1}^{k} p_i f(x|\theta, y_i) = \sum_{j=1}^{k'} p'_j f(x|\theta', y'_j) \quad \text{a.a. } x[\lambda], \tag{18}$$

where $0 < p_i, p'_j < 1$ and $\sum_i p_i = \sum_j p'_j = 1$.

First consider the case where (a) holds for $y < y'$. Then it is convenient to assume (with no loss of generality) that the atoms y_i and y'_j are ordered so that $y_1 < y_2 < \cdots < y_k$ and $y'_1 < y'_2 < \cdots < y'_k$. Furthermore, we can also assume that $y_1 \leq y'_1$. First consider the case if y_1 were strictly smaller than y'_1. The values of $\exp(b(\theta, y_i))$ and $\exp(b(\theta', y'_j))$ remain finite and strictly bounded away from zero over the finite collection of atoms $\{y_i, y'_j\}$. Therefore, if we divide (18) on both sides by $f(x|\theta, y_1)$ and let $x \to x_0$, then the left-hand side equals p_1, while the right-hand side is either 0 or $+\infty$ due to (16). Therefore, we must suppose that (18) holds with $y_1 = y'_1$. Divide (18) by $f(x|\theta, y_1)$ and let $x \to x_0^*$. If $\theta > \theta'$ or $\theta < \theta'$, we get $p_1 = +\infty$ by (16) and (17), otherwise if $\theta = \theta'$ we get $p_1 = p'_1$ by (16). Therefore, it must be that $p_1 = p'_1$, $\theta = \theta'$ and $y_1 = y'_1$.

Now cancel the first term on the left and right-side sums of (18). Apply a similar argument as before on the first term of the new sums. Do this recursively $k - 1$ times, obtaining $\theta = \theta'$ and $(p_i, y_i) = (p'_i, y'_i)$ for $i = 1, \dots, k$. It follows that $k = k'$.

When (a) holds with $y > y'$, order the y_j and y'_j atoms in decreasing order (with no loss of generality). Assume that $y_1 \geq y'_1$ and argue as before that $y_1 > y'_1$ leads to a contradiction. The rest of the proof is the same. □

Theorem 4 can be used to establish the identification in our four motivating examples, as well as the inverse Gaussian mixture.

Example (Normal identification). The conditions for Theorem 4 hold with $x_0 = x_0^* = -\infty$. For $\theta, \theta' > 0$ and $y, y' \in \mathbb{R}$,

$$D(x, \theta, \theta', y, y') = -\frac{x^2}{2}\left(\frac{1}{\theta} - \frac{1}{\theta'}\right) + x\left(\frac{y}{\theta} - \frac{y'}{\theta'}\right). \tag{19}$$

If $\theta = \theta'$ and $y < y'$, then (19) becomes $x(y - y')/\theta$, which converges to $+\infty$ as $x \to -\infty$. If $\theta \neq \theta'$, then (19) equals $-x^2(1/\theta - 1/\theta')/2 + O(x)$, which converges to $+\infty$ as $x \to -\infty$ if $\theta > \theta'$ or $-\infty$ if $\theta < \theta'$.

Example (Negative binomial identification). Use Theorem 4 with $x_0 = x_0^* = +\infty$ and

$$\begin{aligned} &D(x, \theta, \theta', y, y') \\ &\qquad = x(y - y') + \log\left(\frac{\Gamma(x+\theta)}{\Gamma(\theta)}\right) - \log\left(\frac{\Gamma(x+\theta')}{\Gamma(\theta')}\right), \end{aligned} \tag{20}$$

for $\theta, \theta' > 0$ and $y, y' < 0$. By the mean value theorem, the difference in the second and third term in (20) can be expressed as

$$(\theta - \theta')\frac{\partial}{\partial\theta}\log\left(\frac{\Gamma(x+\theta^*)}{\Gamma(\theta^*)}\right),$$

for θ^* between θ and θ'.

By expansion (12), the derivative equals

$$\frac{x}{\theta^*(x+\theta^*)} + \sum_{j=1}^{\infty} \frac{x}{j^2}\left(1+\frac{x+\theta^*}{j}\right)^{-1}\left(1+\frac{\theta^*}{j}\right)^{-1}. \tag{21}$$

Use the monotone convergence theorem to deduce that (21) converges uniformly over θ^* to $+\infty$ at a $o(x)$ rate when $x \to +\infty$. Therefore, 20 equals $x(y-y') + o(x)$, which converges to $+\infty$ as $x \to +\infty$ for $y > y'$. Because (21) is positive, it follows that when $y = y'$ and $\theta < \theta'$, (20) converges to $-\infty$ as $x \to +\infty$. □

EXAMPLE (Gamma identification). For the gamma mixture, use Theorem 4 with $x_0 = x_0^* = +\infty$ and

$$\begin{aligned} &D(x,\theta,\theta',y,y') \\ &\qquad = -x(y-y') + (\theta-\theta')\log x \quad \text{for } \theta,\theta' > 0 \text{ and } y,y' > 0. \end{aligned} \tag{22}$$

When $y < y'$, this equals $-x(y-y') + O(\log x)$, which converges to $+\infty$ as $x \to +\infty$. When $y = y'$, (22) converges to $\pm\infty$ as $x \to +\infty$, depending upon whether $\theta > \theta'$ or not.

EXAMPLE (Weibull identification). Use Theorem 4 with $x_0 = x_0^* = +\infty$. For $\theta, \theta' > 0$ and $y, y' > 0$,

$$D(x,\theta,\theta',y,y') = -(yx^{\theta} - y'x^{\theta'}) + (\theta-\theta')\log x. \tag{23}$$

When $\theta = \theta'$, this becomes $-x^{\theta}(y-y')$ which converges to $+\infty$ as $x \to +\infty$ for $y < y'$. When $\theta \neq \theta'$, write (23) as $-x^{\theta'}(yx^{\theta-\theta'} - y') + O(\log x)$. As $x \to +\infty$ this converges to $+\infty$ for $\theta < \theta'$ and $-\infty$ for $\theta > \theta'$.

EXAMPLE (Inverse Gaussian identification). The inverse Gaussian distribution (sometimes called the Wald distribution), has the density

$$\begin{aligned} f(x|\theta,y) &= \frac{y^{1/2}x^{-3/2}}{\sqrt{2\pi}}\exp\left(-\frac{y(x-\theta)^2}{2\theta^2 x}\right) \\ &= \exp\left(-y\left[\frac{(x-\theta)^2}{2\theta^2 x}\right] - \frac{3}{2}\log x + \frac{1}{2}\log\left(\frac{y}{2\pi}\right)\right), \qquad x > 0, \end{aligned}$$

where $\theta > 0$ acts as a location parameter and $y > 0$ as a dispersion parameter. This is of the form (1) with $\Theta \otimes \mathscr{Y} = \mathbb{R}^+ \otimes \mathbb{R}^+$.

With some work it can be shown that the inverse Gaussian satisfies regularity Condition C1 of Theorem 3. However, the non-$\sqrt{n}$ rate does not necessarily hold here because of the noninvertibility of $s(\cdot,\theta)$. Nevertheless, it is still instructive to study the identification for the model, because unlike the previous examples, its identification follows from Theorem 4 by using different values for x_0 and x_0^* (in this case, $x_0 = 0$ and $x_0^* = +\infty$). To verify the conditions of the theorem, note that for $\theta, \theta' > 0$ and $y, y' > 0$,

$$D(x,\theta,\theta',y,y') = -\frac{1}{2x}\left[\frac{y}{\theta^2}(x-\theta)^2 - \frac{y'}{\theta^2}(x-\theta')^2\right]. \tag{24}$$

For small x, (24) can be written as $(y' - y)/(2x) + O(1)$ which converges to $+\infty$ as $x \to x_0 = 0$ for $y < y'$. For large x, express (24) as $x(y'/(\theta')^2 - y/\theta^2)/2 + O(1)$. If $\theta = \theta'$, this becomes $x(y' - y)/(2\theta^2) + O(1)$, which converges to $+\infty$ as $x \to x_0^* = +\infty$ for $y < y'$. If $\theta < \theta'$ but $y = y'$, (24) becomes $xy(1/(\theta')^2 - 1/\theta^2)/2 + O(1)$, which converges to $-\infty$ as $x \to x_0^* = +\infty$.

APPENDIX

Proof of Lemma 1. Use the $h \in L_0^2(G_0)$ to define h_τ as

$$h_\tau = h\{|h| \le |\tau|^{-1/2}\} - G_0 h\{|h| \le |\tau|^{-1/2}\}.$$

The centering ensures that $G_0 h_\tau = 0$ and the truncation forces $|\tau h_\tau| \le 2|\tau|^{1/2} = o(1)$. Thus, for a small enough τ, the G_τ defined by h_τ through (5) describes a proper distribution because it integrates to one and is positive. (Note: there is nothing special about the $\tau^{-1/2}$ truncation level; it is chosen simply for convenience although other levels work equally well.)

By (i) of Condition C1, we can use the mean value theorem to expand the perturbed density

$$f(x|\theta_0 + \tau, y) = f(x|\theta_0, y) + \tau\Delta(x, y, \theta_0 + \tau^*) \quad \text{for a.a. } (x, y)[\lambda \otimes G_0],$$

where $\tau^* = \tau^*(x, y, \tau)$ and $|\tau^*| \le |\tau|$. Divide throughout by f_0 on the set where it is nonzero (which has P_0 measure 1), and take expectations with respect to G_τ to write

$$\frac{f_\tau(x)}{f_0(x)} = \frac{1}{f_0(x)} \int (f(x|\theta_0, y) + \tau\Delta(x, y, \theta_0 + \tau^*))(1 + \tau h_\tau(y))\, dG_0(y).$$

Expanding, the right-hand side can be written as

$$\begin{aligned}(25)\qquad & 1 + \frac{\tau}{f_0(x)}\left[\int h(y) f(x|\theta_0, y)\, dG_0(y) + \int \Delta(x, y, \theta_0)\, dG_0(y)\right] \\ & + \frac{\tau}{f_0(x)} \int (h_\tau(y) - h(y)) f(x|\theta_0, y)\, dG_0(y) \\ & + \frac{\tau^2}{f_0(x)} \int h_\tau(y)\Delta(x, y, \theta_0 + \tau^*)\, dG_0(y) \\ & + \frac{\tau}{f_0(x)} \int (\Delta(x, y, \theta_0 + \tau^*) - \Delta(x, y, \theta_0))\, dG_0(y).\end{aligned}$$

Recognize that the coefficient of τ in the second term equals $A(G_0, h) + \rho$. Collecting remainder terms,

$$(26)\qquad \frac{f_\tau(x)}{f_0(x)} - 1 = \tau(A(G_0, h)(x) + \rho(x)) + R(x, h, \tau)$$

for a.a. $x[P_0]$.

To prove the lemma, we need to show that $A(G_0, h)$ and ρ are $L_0^2(P_0)$ functions, and that $R(\cdot, h, \tau)$ has a squared $L^2(P_0)$ norm $o(\tau^2)$. A convenient bound for this task is as follows. For $\lambda \otimes G_0$-measurable ψ,

$$\begin{aligned} P_0\left(\frac{G_0\psi}{f_0(x)}\right)^2 &= \int \frac{1}{f_0(x)}\left[\int \frac{\psi(x,y)}{f(x|\theta_0,y)^{1/2}} f(x|\theta_0,y)^{1/2}\, dG_0(y)\right]^2 d\lambda(x) \\ &\le \iint \frac{\psi(x,y)^2}{f(x|\theta_0,y)}\, dG_0(y)\, d\lambda(x), \end{aligned} \tag{27}$$

where the last inequality is a result of the Cauchy–Schwarz inequality.

Start with the third term in (25). Use (27) to bound its squared $L^2(P_0)$ norm by

$$\tau^2 \iint (h_\tau(y) - h(y))^2 f(x|\theta_0,y)\, dG_0(y)\, d\lambda(x) = \tau^2 G_0(h_\tau - h)^2.$$

Because $h_\tau \to h$, this term must be $o(\tau^2)$ by the dominated convergence theorem ($|h_\tau|$ can be bounded by $|h| + G_0|h|$).

Bound h_τ by $2|\tau|^{-1/2}$ and use (27) to bound the squared P_0 expectation of the fourth term in (25) by

$$4|\tau|^2 \iint \frac{\Delta(x,y,\theta_0+\tau^*)^2}{f(x|\theta_0,y)}\, dG_0(y)\, d\lambda(x).$$

Use (iii) of Condition C1 to deduce that this term is $O(\tau^3)$.

Now consider the last term in (25). Using (27), its squared P_0 expectation is bounded by

$$\tau^2 \iint \frac{1}{f(x|\theta_0,y)} [\Delta(x,y,\theta_0+\tau^*) - \Delta(x,y,\theta_0)]^2\, dG_0(y)\, d\lambda(x).$$

Use the dominated convergence theorem with dominating function $4M$ and the continuity of Δ in τ ((ii) of condition C1) to deduce that this term is $o(\tau^2)$.

Similar arguments using (27) show that both $A(G_0, h)$ and ρ are in $L^2(P_0)$ [you need (iii) of condition C1 for the ρ part]. Therefore, if we square (26) and take its P_0 expectation we arrive at (10). To complete the proof we need to show that ρ and $A(G_0, h)$ have zero P_0 expectation. The proof for $A(G_0, h)$ is straightforward. For ρ, take the P_0 expectation of (26). From $P_0 A(G_0, h) = 0$ and $P_0 R(\cdot, h, \tau) = o(\tau)$, deduce that $P_0 \rho = 0$. □

PROOF OF LEMMA 2. Suppose $\phi \in L_0^2(P_0)$ such that $P_0 \phi A(G_0, h) = 0$ for each $h \in L_0^2(G_0)$. To prove that the range of $A(G_0, \cdot\,)$ is dense in $L_0^2(P_0)$, we will show that $\phi = 0$ a.e. $[P_0]$.

Interchanging the order of integration [both ϕ and $A(G_0, h)$ are elements of $L_0^2(P_0)$], we see that ϕ must satisfy

$$\iint h(y)\phi(x) \frac{f(x|\theta_0,y)}{f_0(x)}\, dP_0(x)\, dG_0(y) = 0 \quad \text{for each } h \in L_0^2(G_0).$$

That is, $G_0 hT = 0$ for each $h \in L_0^2(G_0)$, where

$$T(y) = \int \phi(x) f(x|\theta_0, y)\, d\lambda(x).$$

Using the Cauchy–Schwarz inequality,

$$G_0 T^2 \le G_0 \int \phi(x)^2 f(x|\theta_0, y)\, d\lambda(x) = P_0 \phi^2 < \infty.$$

Furthermore, $G_0 T = P_0 \phi = 0$, so that $T \in L_0^2(G_0)$. Therefore, if we choose $h = T$, we have that $G_0 hT = G_0 T^2 = 0$. In particular, this implies that $T = 0$ over the support for G_0.

Our assumption of a limit point for G_0 ensures that we can find a sequence $y_j \in \mathscr{Y}$ converging to an interior point $y_0 \in \mathscr{Y}$ such that $T(y_j) = 0$. Dividing by positive $b(\theta_0, y_j)$, the equality $T(y_j) = 0$ implies that $g(y_j) = 0$, where

$$g(y) = \int \phi(x) \exp(ys(x, \theta_0))\, d\lambda^*(x)$$

and $d\lambda^*(x) = \exp(t(x, \theta_0))\, d\lambda(x)$.

Use the Cauchy–Schwarz inequality to deduce that

$$\begin{aligned} &\left(\int |\phi(x)| \exp(ys(x, \theta_0) + t(x, \theta_0))\, d\lambda(x) \right)^2 \\ &\quad \le \int \phi(x)^2 f_0(x)\, d\lambda(x) \int \frac{\exp(2ys(x, \theta_0) + 2t(x, \theta_0))}{f_0(x)}\, d\lambda(x) \qquad (28) \\ &\quad = P_0 \phi^2 \times \int \left(G_0 \exp[(Y - 2y)s(x, \theta_0) + b(\theta_0, Y)] \right)^{-1} d\lambda^*(x). \end{aligned}$$

By showing that the right-hand side is finite over some interval containing y_0, we will be able to utilize a Fourier argument which shows that $\phi = 0$ a.e. $[P_0]$. By definition, $\phi \in L_0^2(P_0)$. Thus, to show finiteness, bound the second integral by restricting the range of integration of G_0 to the set $\mathscr{Y}(\varepsilon) = [y_0 - \varepsilon, y_0 + \varepsilon]$, where $\varepsilon > 0$ is chosen small enough so that $\mathscr{Y}(\varepsilon) \subseteq \mathscr{Y}$. For all $y, y' \in \mathscr{Y}(\varepsilon)$,

$$\begin{aligned} \exp((2y' - y)s(x, \theta_0)) &\le \exp((y_0 + 3\varepsilon)s(x, \theta_0) \vee (y_0 - 3\varepsilon)s(x, \theta_0)) \\ &\le \exp((y_0 + 3\varepsilon)s(x, \theta_0)) + \exp((y_0 - 3\varepsilon)s(x, \theta_0)). \end{aligned}$$

Use this to bound the second integral on the right of (28) by

$$\begin{aligned} &[G_0\{Y \in \mathscr{Y}(\varepsilon)\} \exp(b(\theta_0, Y))]^{-1} \\ &\qquad \times [\exp(-b(\theta_0, y_0 + 3\varepsilon)) + \exp(-b(\theta_0, y_0 - 3\varepsilon))]. \end{aligned}$$

By continuity, $\exp(-b(\theta_0, y))$ achieves a finite maxima over $\mathscr{Y}(3\varepsilon)$, where we can assume ε is small enough so that $\mathscr{Y}(3\varepsilon) \subseteq \mathscr{Y}$. Therefore, the expression above is finite from the fact that $G_0\{Y \in \mathscr{Y}(\varepsilon)\} > 0$. Consequently, (28) implies that for a small enough $\varepsilon > 0$,

$$(29) \qquad \int |\phi(x)| \exp(ys(x, \theta_0))\, d\lambda^*(x) < \infty \quad \text{for } y \in \mathscr{Y}(\varepsilon).$$

To establish that $\phi = 0$, use a standard result concerning exponential families to deduce that the finiteness of (29) implies that g is analytic over $\{z: \Re(z) \in \mathscr{Y}(\varepsilon)\}$. Because $g(y_j) = 0$ over the sequence $y_j \to y_0$, we find by analytic continuation that g must be identically zero over its region of analyticity. Define $\phi_s(u) = \phi(s^{-1}(u, \theta_0))$, where s^{-1} is the inverse of s. Then, ϕ_s is measurable (by the measurability of s^{-1}) over the σ-algebra $\mathscr{S}$ induced by $s(\cdot, \theta_0)$. Make the change of variables $s = s(x, \theta_0)$ in the integrand of g. Then for each $y' \in \mathscr{Y}(\varepsilon)$,

$$\int \phi_s(s) \exp(its)\, d\lambda'_s(s) = 0 \quad \text{for } t \in \mathbb{R},$$

where λ'_s is the finite (image) measure defined by

$$\lambda'_s(B) = \int \{x: s(x, \theta_0) \in B\} \exp(y's(x, \theta_0) + t(x, \theta_0))\, d\lambda(x),$$

for each set $B \in \mathscr{S}$.

Use a standard Fourier transform argument by breaking ϕ_s into its positive and negative components to deduce that $\phi_s = 0$ a.e. $[\lambda'_s]$. It follows from the bimeasurability of $s(\cdot, \theta_0)$ that $\phi = 0$ a.e. $[\lambda]$ and therefore that $\phi = 0$ a.e. $[P_0]$. □

Acknowledgments. The author is grateful to an Associate Editor and referee who helped review a first draft of this paper. In particular, the Associate Editor's comments were extremely helpful in restructuring the paper so as to bring out key ideas more clearly.

REFERENCES

AHLFORS, L. V. (1979). *Complex Analysis*, 3rd ed. McGraw-Hill, New York.

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.

CHEN, J. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233.

ELBERS, C. and RIDDER, G. (1982). True and spurious duration dependence: the identifiability of the proportional hazard model. *Rev. Econom. Stud.* **49** 403–410.

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.

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

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

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27** 887–906.

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

LINDSAY, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11** 486–497.

LINDSAY, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. IMS, Hayward, CA.

TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **32** 1265–1269.

VAN DER VAART, A. W. (1988). Estimating a real parameter in a class of semiparametric models. *Ann. Statist* **16** 1450–1474.

DEPARTMENT OF BIOSTATISTICS
AND EPIDEMIOLOGY—WB4
CLEVELAND CLINIC FOUNDATION
9500 EUCLID AVENUE
CLEVELAND, OHIO 44195