Columbia University

Department of Economics
Discussion Paper Series

Identifying Finite Mixtures in Econometric Models

Marc Henry
Yuichi Kitamura
Bernard Salanié

Discussion Paper No.: 0910-20

Department of Economics
Columbia University
New York, NY 10027

August 2010

Identifying Finite Mixtures in Econometric Models

Marc Henry[1], **Yuichi Kitamura**[2] **and Bernard Salanié**[3]

August 10, 2010.

[1]Département de sciences économiques, Université de Montréal, CIRANO, CIREQ. E-mail: marc.henry@umontreal.ca. Parts of this paper were written while Henry was visiting the University of Tokyo Graduate School of Economics; he gratefully acknowledges the CIRJE for its support.

[2]Economics Department, Yale University. E-mail: yuichi.kitamura@yale.edu. Kitamura gratefully acknowledges the NSF (grants SES-0551271 and SES-0851759) for its support.

[3]Economics Department, Columbia University. E-mail: bs2237@columbia.edu. Parts of this paper were written while Salanié was visiting the Toulouse School of Economics; he gratefully acknowledges the Georges Meyer endowment for its support.

Abstract

Mixtures of distributions are present in many econometric models, such as models with unobserved heterogeneity. It is thus crucial to have a general approach to identify them nonparametrically. Yet the literature so far only contains isolated examples, applied to specific models. We derive the identifying implications of a conditional independence assumption in finite mixture models. It applies for instance to models with unobserved heterogeneity, regime switching models, and models with mismeasured discrete regressors. Under this assumption, we derive sharp bounds on the mixture weights and components. For models with two mixture components, we show that if in addition the components behave differently in the tails of their distributions, then components and weights are fully nonparametrically identified. We apply our findings to the nonparametric identification and estimation of outcome distributions with a misclassified binary regressor. This provides a simple estimator that does not require instrumental variables, auxiliary data, symmetric error distributions or other shape restrictions.

Introduction

Many models used in econometrics can be interpreted as mixtures of distributions. Models with unobserved heterogeneity are an obvious example (see, for example, Heckman and Singer (1984) and Cameron and Heckman (1998) for applications in labor economics). Many dynamic structural models use mixtures to incorporate unobserved heterogeneity (e.g. Keane and Wolpin (1997)). Further examples of models with unobserved heterogeneity are given in Section 2.

As is well known, regime switching models also are mixture models. Take the Hamilton (1989) model: we have

$$y_t = a(s_t) + \varepsilon_t$$

where the mean term $a(\cdot)$ depends on the state s_t, the innovation ε_t is independent from the past values $y_1, ..., y_{t-1}$ of y and s_t is a hidden Markov chain taking values 0 or 1. The density of y_t conditional on its history is, letting f denote the density of ε, given by

$$f(y_t - a(s_t))$$

if we knew the realization of the Markov chain $(s_1, \ldots, s_T)$. As we do not, we need to integrate over its possible values. The conditional density of y_t given its past values becomes

$$\sum_{s=0,1} f(y_t - a(s_t)) \Pr(s_t = s | y_1, \ldots, y_{t-1})$$

which is recognizable as a mixture. Similarly, stochastic volatility models (e.g. Kim and Nelson (1998))

$$y_t = \exp(v_t/2)\varepsilon_t$$

with unobserved volatility

$$v_t = a + bv_{t-1} + \sigma u_t$$

belong to the family of hidden Markov chains, which are all mixture models.

Measurement error models with discrete regressors are finite mixture models (see the surveys by Carroll, Ruppert, Stefanski, and Crainiceanu (2006) and Chen, Hong, and Nekipelov (2009)) as are models with data contamination (Horowitz and Manski (1995)).

As shown by these examples, a large number of models commonly used in econometrics are conditional mixture models: they explain the distribution, or some moments, of a random variable y through a mixture of statistical models that are conditional on another

random variable z[1]. A conditional mixture model is written as

$$F_{Y|Z}(y|z) = \int F_{Y|Z,S}(y|z,s)F_{S|Z}(ds|z),$$

where the F notation denotes a cdf. We call s the *mixture index*, $F_{S|Z}(ds|z)$ the *mixture weights*, and each $F_{Y|Z,S}(.|.;s)$ a *mixture component.*.

Our main focus in this paper will be on the identification of $F_{Y|S,Z}$ and $F_{S|Z}$ when $F_{Y|Z}$ is directly identified from the data. If the object of interest is not a cdf F but a linear functional T_F of it, all of our results would apply without any change—the case in which T_F is $E_F G$ for a given function G of the random variables of course is a case in point.

Without further assumptions there is of course no way to identify the mixture weights and components: e.g. choosing a mixture index t and putting all the weight on component t with $F_{Y|Z,S}(y|z;t) \equiv F_{Y|Z}(y|z)$ rationalizes the data.

The case when the distribution of the mixture index has finite support has generated a voluminous literature in statistics. The main focus of the literature (see e.g. Teicher (1963), Farewell (1982) Lindsay and Roeder (1993)) has been on parametric identification and on two difficult problems: numerical instabilities in estimation, and testing for the number of mixture components. Some recent work has started to come to terms with nonparametric identification of mixture models, continuous or finite. This literature uses several approaches, depending on the specific problem it attempts to solve.

- In a regression model with mismeasured binary regressor model with conditional independence, Mahajan (2006) and Lewbel (2007) have proposed an instrumental variable approach; Chen, Hu, and Lewbel (2008b) and Chen, Hu, and Lewbel (2009) rely on moment restrictions on the error term.

- Hu (2008) uses a similar strategy in discrete regressor models in the presence of nonlinear measurement error, while Chen, Hu, and Lewbel (2008a) again impose moment restrictions on the error term.

- Hall and Zhou (2003) and Kitamura (2003) study nonparametric identification of finite mixtures without relying on exclusion restrictions. An interesting related paper on this subject is Kasahara and Shimotsu (2008), which, as in Kitamura (2003), uses identification power of covariates, and then applies the technique developed by Hall and Zhou (2003). Their result is useful in analyzing conditional choice probability

[1]The term "random variable" here should not be taken to imply that we focus on scalar-valued variables: y and z can take values in a vector space, for instance.

(CCP) type estimation procedures, such as Hotz and Miller (1993), which have gained renewed attention in the recent literature on dynamic structural estimation. In a related application, Hu and Shum (2008) consider identification of a Markov chain with some unobserved components.

Our first contribution is to set up a general identification strategy in finite mixture models; we show how a single exclusion restriction that underlies most previous work generates partial identification without further assumptions. This exclusion restriction consists in assuming that each component of the mixture $F_{Y|Z,S}(y|z;s)$ only depends on some components of the conditioning random variable z.

To be more precise, we assume that $z = (x', w')'$ and that for all s and z,

$$F_{Y|Z,S}(.|z;s) \equiv F_{Y|X,S}(.|x;s)$$

is independent of w conditional on x and s. This exclusion restriction needs to be supplemented with only one more assumption: that the mixture weights do depend on w conditional on x. We state these two assumptions as Assumptions 1 and 2 in section 1.

The surprising fact, proved in section 1, is that in finite mixture models we do not actually need much more: these two assumptions identify the mixture weights and the mixture components nonparametrically, up to a linear transformation whose coefficients can only be functions of x. Moreover, these coefficients must satisfy a simple set of linear inequalities; these describe a partially identified region for the parameters of interest. For notational simplicity, we then focus on the case when only two components coexist (s can only take two values.) We first discuss our exclusion restriction in section 2; we show that it holds by construction in a number of the models that have been used in the literature. We give a detailed study of the partially identified region in section 3, and we show that some quantities of interest are actually point identified from our two assumptions.

An interesting feature of our approach is that Assumption 1 has testable consequences: it implies a multiplicative separability property that can be checked on the estimates, in the manner of an overidentification test (even though the model itself is only partially identified.)

It is actually possible in many applications to go from partial identification to full nonparametric identification of the components and weights. To do this, we rely on a argument that is a rather natural version of "identification at infinity". Suppose individuals in a population are indexed by an unobserved binary characteristic, say "high" and "low". If the high value of the index tends to correspond to relatively good outcomes, then upon

observing a very good outcome the analyst would normally find it quite likely that the individual belongs to the population with the high index value. We formalize and extend this intuition as Assumption 3 in section 4 and we show that if two such restrictions hold (say the one above and a polar one for very bad outcomes), then in the binary mixture model components and weights are nonparametrically point identified.

This suggests a fairly simple fully nonparametric estimation procedure; we examine its properties in the misclassified binary regressor model in section 5, and we appeal to results on the tail empirical process to prove that it yields a consistent and asymptotically normal estimator. The identification strategy has important similarities with traditional identification at infinity, as in Chamberlain (1986) and Heckman (1990). The identification relies on the fact that the propensity score tends to one in the right tail and to zero in the left tail of the distribution of outcomes. However, it relies on the distributions of outcomes themselves and not on the distribution of covariates, and thus is more straightforward and easier to rationalize. While our estimator is asymptotically normal, it converges more slowly than the parametric rate.

The paper is organized as follows. Section 1 presents general partial identification results for an arbitrary finite mixture of distributions. Section 2 develops three major classes of econometric applications of our framework in the case of mixtures of two component distributions. Section 3 develops sharp bounds on the mixture weights and component distributions, with special emphasis on the case of mismeasured binary regressors and possibly missing observations. Section 4 gives the identification results based on relative tail behaviour of the component distributions, and section 5 applies the identification rationale to the construction of a very simple and asymptotically normal estimator of causal effect distributions. The last section concludes.

1 Partial Identification of Finite Mixtures

Let y and z be two random variables; we assume that the conditional model is a mixture of $J > 1$ otherwise unrestricted components:

$$F(y|z) = \sum_{j=0}^{J-1} \lambda_j(z) F_j(y|z).$$

The mixture weights $\lambda_j(z)$ are non-negative and sum to 1; we allow for the possibility that some of them are actually zero, so that the model has fewer than J components for some or all values of z. The assumptions and results are stated for the case where all weights are

positive, as this simplifies the statements.

As explained in the introduction, our main assumption is an exclusion restriction: the conditioning variables $z = (x', w')'$ are such that

Assumption 1 *The variable w is excluded from the component distributions' conditioning set, i.e.*

$$F_j(y|z) = F_j(y|x),$$

for all $j = 0, \ldots, J-1$ and all possible values of (y, z).

We put off the discussion of Assumption 1 to section 2; for now, we focus on proving that it is sufficient to ensure that the J components of the mixture and their $(J-1)$ weights are identified up to $J(J-1)$ functions of x.

Take the first mixture component as reference point and define the new unknown quantities

$$\begin{aligned} \Delta(y|x) &= (F_1(y|x) - F_0(y|x), \ldots, F_{J-1}(y|x) - F_0(y|x))' \\ \text{and } \lambda(z) &= (\lambda_1(z), \ldots, \lambda_{J-1}(z)). \end{aligned}$$

Then we can write

$$F(y|z) = F_0(y|x) + \Delta(y|x)'\lambda(w, x);$$

and if w and w' are two values of the excluded variable,

$$F(y|w, x) - F(y|w', x) = \Delta(y|x)'(\lambda(w, x) - \lambda(w', x)). \qquad (M)$$

This equation is the basis for our identification strategy. One very strong implication is worth noting: for given x, the function of three variables

$$(y, w, w') \longrightarrow F(y|w, x) - F(y|w', x)$$

is a scalar product of two $(J-1)$-dimensional vectors; the first one is only a function of y, and the other one is an additively separable, antisymmetric function of w and w' only. This points towards overidentification tests, as mentioned before.

The case $J = 2$ is of special interest to us, and it makes the testable implications of the exclusion restriction even more transparent: now conditional on x,

$$F(y|w, x) - F(y|w', x) = (\lambda(w, x) - \lambda(w', x))(F_1(y|x) - F_0(y|x))$$

is the product of a scalar function of y and an additively separable, antisymmetric function of w and w' only.

Going back to the general finite case, we define a first "regular" set of values of x:

Definition 1 *Let $\tilde{\mathcal{X}}$ be the set of x such that for some $(w_0, \ldots, w_{J-1})$ in the support of w given x and some $(y_1, \ldots, y_{J-1})$ in the support of y given x, the $(J-1) \times (J-1)$ matrix Δ with (i,j)-th element $F_j(y_i) - F_0(y_i)$ and the $(J-1) \times (J-1)$ matrix Λ with (i,j)-th element $\lambda_i(w_j) - \lambda_i(w_0)$ are invertible.*

Note that the set $\tilde{\mathcal{X}}$ may well be empty. This could only happen in uninteresting cases if $J = 2$, but for larger J it may be for instance that w only takes fewer than J distinct values for all x. Then a fortiori the linear independence property in Definition 1 would fail. Our second assumption rules out such cases, in which the excluded variable w does not generate enough variation in the y variable:

Assumption 2 *The set $\tilde{\mathcal{X}}$ is non-empty.*

Under Assumptions 1 and 2, our model has at least $J(J-1)$ dimensions of indeterminacy. To see this, let $v(x)$ be any $(J-1)$-dimensional vector function of x and $M(x)$ be any function of x whose values are invertible $(J-1)$-dimensional matrices. Start from the true DGP (λ, Δ, F_0) and define

$$\begin{aligned}
\mu(w,x) &= M(x)(\lambda(w,x) + v(x)) \\
\delta(y|x) &= (M(x)^{-1})' \Delta(y|x) \\
G_0(y|x) &= F_0(y|x) - \Delta(y|x)' v(x).
\end{aligned}$$

It is easy to check that the new vector of functions (μ, δ, G_0) also generates the observed cdf $F(y|z)$. The only caveat is that mixture components must remain cdfs and mixture weights must remain probabilities, which imposes linear inequalities on admissible transformations $(v(x), M(x))$. We call a pair (v, M) that satisfies these constraints an admissible (v, M) transform. Note that it has $(J-1) + (J-1)^2 = J(J-1)$ degrees of freedom, subject to linear inequalities.

If $J = 2$, things are much simpler: an admissible (v, M) transform is simply given by a pair of numbers $(v(x), M(x))$ such that $M(x) \neq 0$ and some other linear inequalities hold—these are presented in much more detail in section 3.

Our main result shows that admissible (v, M) transforms in fact exactly define partial nonparametric identification:

Theorem 1 *If assumptions 1 and 2 hold, then for each $x \in \tilde{\mathcal{X}}$, the mixture components $F_0, \ldots, F_{J-1}$ and the mixture weights λ are nonparametrically identified up to an admissible (v, M) transform.*

Proof Take an $x \in \tilde{\mathcal{X}}$ and fix any of the J values $(w_0, \ldots, w_{J-1})$ of w and $(J-1)$ values $(y_1, \ldots, y_{J-1})$ of y that make x an element of $\tilde{\mathcal{X}}$. Drop x from the notation from now on.

Suppose that F and G are observationally equivalent and both satisfy assumptions 1 and 2. Hence we have the following for all y and w.

$$\begin{aligned} F(y|w) &= F_0(y) + \Delta(y)^t \lambda(w), \\ G(y|w) &= G_0(y) + \delta(y)^t \mu(w), \end{aligned}$$

from which we deduce

$$\begin{aligned} F(y|w) - F(y|w_0) &= \Delta(y)^t [\lambda(w) - \lambda(w_0)], \\ G(y|w) - G(y|w_0) &= \delta(y)^t [\mu(w) - \mu(w_0)]. \end{aligned}$$

Hence

$$\Delta(y)^t [\lambda(w) - \lambda(w_0)] = \delta(y)^t [\mu(w) - \mu(w_0)], \tag{1.1}$$

and

$$\Delta[\lambda(w) - \lambda(w_0)] = \delta[\mu(w) - \mu(w_0)],$$

with both Δ and δ invertible. We therefore have

$$\mu(w) = M[\lambda(w) + v],$$

with $M = \delta^{-1}\Delta$ and $v = \Delta^{-1}\delta\mu(w_0) - \lambda(w_J)$, so that equation (1.1) becomes

$$\Delta(y)[\lambda(w) - \lambda(w_0)] = \delta(y) M [\lambda(w) - \lambda(w_0)]$$

which is true for $w = w_j$, all $j = 1, \ldots, J-1$. But since the matrix Λ is invertible, we can conclude that $\delta(y) = M^{-1}\Delta(y)$.

Finally, since F and G are observationally equivalent, we have

$$F_0(y) + \Delta(y)^t \lambda(w) = G_0(y) + \delta(y)^t \mu(w),$$

hence

$$\begin{aligned} G_0(y) &= F_0(y) + \Delta(y)^t \lambda(w) - \Delta(y)^t M^{-1} M [\lambda(w) + v] \\ &= F_0(y) - \Delta(y)^t v, \end{aligned}$$

hence the result. ■

Assumption 2 bears on the true model, and thus cannot be tested directly. We now define a second set of "regular" values of x that only involves observable quantities.

Definition 2 *Let $\mathcal{X}$ be the set of x such that for some $(w_0, \ldots, w_{J-1})$ in the support of w given x and some $(y_1, \ldots, y_{J-1})$ in the support of y given x, the $(J-1) \times J$ matrix with (i,j)-th element $(F(y_i|w_{j-1}, x)$ has full rank.*

Again, this is easier to understand when $J = 2$. Then $x \in \mathcal{X}$ iff there exist w_1 and w_2 that are in the support of $w|x$ and such that $F(.|w_1, x)$ and $F(.|w_2, x)$ do not coincide. Equivalently, we require that for some value of y, the function $w \longrightarrow F(y|w, x)$ takes at least two different values.

Definition 2 is just a generalization to $J > 2$. Take $J = 3$ for instance; then it is easy to see that $x \in \mathcal{X}$ requires the existence of (w_0, w_1, w_2) and (y_1, y_2) that satisfy:

1. $F(y_1|w_2, x) \neq F(y_1|w_0, x)$
2. $F(y_2|w_2, x) \neq F(y_2|w_1, x)$
3. and finally,

$$\frac{F(y_1|w_1, x) - F(y_1|w_2, x)}{F(y_1|w_0, x) - F(y_1|w_2, x)} \neq \frac{F(y_2|w_1, x) - F(y_2|w_2, x)}{F(y_2|w_0, x) - F(y_2|w_2, x)}.$$

Item 1 (resp. 2) simply requires that there exist a value y_1 (resp. y_2) where $F(.|w_0, x)$ (resp. $F(.|w_1, x)$) and $F(.|w_2, x)$ differ. Item 3 is slightly more complex, but in essence it requires that when w moves from w_2 to w_0 or to w_1, the ratios of the changes in the cdf of y are different in y_1 and y_2. It is hard to find non-degenerate examples in which item 3 would not apply[2]

Note that definition 2 involves properties of the observed mixture. It will allow us to state necessary conditions for identification that involve only "observables" (i.e. $F(y|z)$ which is directly identified from the data.) This follows from the following Lemma:

Lemma 1 *If assumptions 1 and 2 hold, then $\tilde{\mathcal{X}} \subseteq \mathcal{X}$.*

Proof Take an $x \in \tilde{\mathcal{X}}$ and fix any of the J values $(w_1, \ldots, w_J)$ of w and $J-1$ values $(y_1, \ldots, y_{J-1})$ of y that make x an element of $\mathcal{X}$. From now on, we drop x from the notation.

[2]Uniform distributions of y on a fixed-length interval whose location depends on w seem to be pretty much the only possibility.

By assumption 2, the $(J-1)\times(J-1)$ matrix A with (i,j)-th element $A_{ij} = F(y_i|w_j) - F(y_i|w_J)$ is invertible. If it were otherwise, then we would have

$$\sum_{j=1}^{J-1} A_{ij} u_j = 0$$

for all i and some nonzero vector $(u_1, \ldots, u_{J-1})$; but this gives

$$\sum_{j=1}^{J-1} F(y_i|w_j) u_j = F(y_i|w_J) \sum_{j=1}^{J-1} u_j$$

which contradicts $x \in \tilde{\mathcal{X}}$.

By assumption 1, the matrix A is the product $\Delta \times \Lambda$, where Δ is the $(J-1)\times(J-1)$ matrix with (i,j)-th element $F_j(y_i) - F_J(y_i)$ and Λ is the $(J-1)\times(J-1)$ matrix with (i,j)-th element $\lambda_i(w_j) - \lambda_i(w_J)$. Hence Δ is invertible, and so is Λ. Hence x is in $\mathcal{X}$. ■

Theorem 1 and Lemma 1 immediately imply that

Corollary 1 *If assumptions 1 and 2 hold, then for each $x \in \mathcal{X}$ the mixture components $F_0, \ldots, F_{J-1}$ and the mixture weights λ are nonparametrically identified up to an admissible (v, M) transform.*

This series of results calls for a couple of remarks. First, the intuition for this order of indeterminacy is fairly simple. Fix one particular value of x. Under assumption 1, the distribution of y given w is a sum of products of J functions of y with weights that only depend on w. Clearly, one can apply a transformation matrix to any particular solution in a J-dimensional space. Any such matrix has J^2 elements; but it must also keep the total mass of the weights equal to one, which introduces J restrictions.

Second, the definition of $\mathcal{X}$ only refers to observable quantities: the conditional cdfs $F(y|z)$. Thus in principle it is possible to determine whether a given x actually belongs to $\mathcal{X}$, or at least to get a forewarning of difficulties if the linear independence condition in Definition 2 fails.

Finally and as already mentioned, assumption 1 generates overidentifying restrictions. Thus even though the model is only partially identified under assumptions 1 and 2, it is still possible to test and reject either of these assumptions.

2 Mixtures with Two Components

From now on we focus on mixtures of two components, i.e. $J = 2$. We do this for several reasons: first, several of the main applications that we mentioned in the introduction explicitly assume two components. Second, it simplifies our notation, thereby easing our exposition of identification results. Third, the order of indeterminacy $J(J-1) = 2$ is quite manageable with only two components; but with more components any attempt to impose further restrictions must be much more model-specific, so that there is less scope for a general discussion.

With two mixture components, Assumption 1 becomes

$$\begin{aligned} F(y|z) &= \lambda(z)F_1(y|x) + (1-\lambda(z))F_0(y|x) \\ &= \lambda(z)\Delta(y|x) + F_0(y|x), \end{aligned}$$

where $\Delta(y|x) = F_1(y|x) - F_0(y|x)$.

As explained in section 1, with two components $\mathcal{X}$ is simply the set of values x such that the function $w \longrightarrow F(.|w,x)$ takes at least two values (in the space of cdfs.) Take such a value $x \in \mathcal{X}$; then it follows from corollary 1 that in x, the model is identified up to two numbers.

We now review three classes of applications and we show that assumption 1 holds in each case. In later sections, we shall illustrate our partial and point identification results in each of these models.

2.1 Models with misclassified binary regressor

Consider a regression model with misclassified binary regressor, where y is the regression outcome. The true regressor $T^* = 0, 1$ is unobserved by the econometrician, who only observes reported status $T = 0, 1$. In addition, even this may be missing for some individuals; we then write $T = \varnothing$.

We drop other observable covariates from the notation, but they could be incorporated with trivial changes. We have the following identity:

$$F(y|T) = \sum_{s=0,1} F(y|T^* = s, T) \Pr(T^* = s|T).$$

In this application $z = T$. The variable w that we exclude in assumption 1 can only be T, so that

$$F(y|T^*, T) \equiv F(y|T^*).$$

Thus we require that the cdf of outcomes for any group with actual regressor value T^* does not depend on the reported value. This assumption is imposed in recent work on identification with misclassified binary regressors, including Mahajan (2006), Hu (2008), Lewbel (2007) and Chen, Hu, and Lewbel (2009), as surveyed in Chen, Hong, and Nekipelov (2009). It should hold if errors in regressor values are due to clerical mistakes; on the other hand, we would expect it to fail if individuals can manipulate reports and causal effects vary across observationally identical individuals.

Given this exclusion restriction, the model becomes

$$F(y|T) = \sum_{s=0,1} F(y|T^* = s) \Pr(T^* = s|T),$$

where the components to be identified are the cdf $F(y|T^* = 1)$ of outcomes when $T^* = 1$, the cdf $F(y|T^* = 0)$ of outcomes when $T^* = 0$, and the probabilities of $T^* = 1$ given information $\Pr(T^* = 1|T)$ for $T = 0, 1, \varnothing$. Assumption 2 requires that $\Pr(T^* = s|T)$ depend on T. This holds automatically if report (or missing report, as we will see later) is informative on the actual value, as one would expect.

The identification strategy in Mahajan (2006), Hu (2006) and Lewbel (2007) relies on an additional instrument, whereas Chen, Hu, and Lewbel (2009) rely on a moment condition on the measurement error[3]. In both cases, only results on expectations are provided, whereas we give here results on the distributions of outcomes. We provide partial identification results, and we we show that missing data can be informative: in many cases it helps shrink the size of the identified regions. Moreover, we prove nonparametric point identification with a strategy of identification at infinity. Our main results are concerned with the identification of $F(y|T^* = s), s = 0, 1$ and $\Pr(T^* = s|T = s'), s, s' = 0, 1$. If we further let $y = T^* y_1 + (1 - T^*)y_0$ and impose the standard unconfoundedness assumption, that is, $(y_0, y_1) \perp\!\!\!\perp T^*$ (possibly conditional on covariates), then it is immediate that the average treatment effect $E[y_1] - E[y_0]$ is identified in the presence of misclassification in treatments.

2.2 Regime switching

Consider the Markov switching model discussed in the introduction, where y_t, $t = 1, \ldots, T$ is independently and identically distributed conditionally on a state variable $S_t \in \{0, 1\}$ that follows a Markov chain with transition probabilities

$$\begin{aligned} \Pr(S_t = 1|S_{t-1} = 1) &= P_{11} \\ \Pr(S_t = 0|S_{t-1} = 0) &= P_{00}. \end{aligned}$$

[3]Their condition holds in particular if measurement error is symmetric.

In such a model, Assumption 1 is automatically satisfied with $y = y_t$ and $w = y_{t-1}$. Indeed, we have

$$F(y_t|y_{t-1} = w) = \lambda(w)F(y_t|S_t = 1) + (1 - \lambda(w))F(y_t|S_t = 0).$$

We can also get a simple closed form formula for the mixture weight $\lambda(w)$ if the mixture components are known; this can be very useful in applications.

Lemma 2

$$\lambda(w) := P(S_t = 1|y_{t-1} = w) = 1 - P_{00} + \frac{P_{11} + P_{00} - 1}{1 + \frac{1 - P_{11}}{1 - P_{00}} \frac{f_0(w)}{f_1(w)}}$$

where f_0 (resp. f_1) is the pdf of y_t conditional on $S_t = 0$ (resp. $S_t = 1$.)

Proof of Lemma 2:

$$\begin{aligned}
&P(S_t = 1 \text{ and } y_{t-1} \leq w) \\
&\quad = P(S_t = 1 \text{ and } y_{t-1} \leq w|S_{t-1} = 1)P(S_{t-1} = 1) \\
&\qquad\qquad + P(S_t = 1 \text{ and } y_{t-1} \leq w|S_{t-1} = 0)P(S_{t-1} = 0) \\
&\quad = P(S_t = 1|S_{t-1} = 1)P(S_{t-1} = 1)P(y_{t-1} \leq w|S_{t-1} = 1) \\
&\qquad\qquad + P(S_t = 1|S_{t-1} = 0)P(S_{t-1} = 0)P(y_{t-1} \leq w|S_{t-1} = 0) \\
&\quad = P_{11}P(S_{t-1} = 1)F_1(w) + (1 - P_{00})P(S_{t-1} = 0)F_0(w).
\end{aligned}$$

(The second equality above uses the exclusion restriction: conditionally on S_{t-1}, y_{t-1} and S_t are independent.)

Moreover,

$$\begin{aligned}
P(y_{t-1} \leq w) &= P(y_{t-1} \leq w|S_{t-1} = 1)P(S_{t-1} = 1) + P(y_{t-1} \leq w|S_{t-1} = 0)P(S_{t-1} = 0) \\
&= P(S_{t-1} = 1)F_1(w) + P(S_{t-1} = 0)F_0(w).
\end{aligned}$$

In addition, the steady state probabilities of the Markov chain are

$$\begin{aligned}
P(S_{t-1} = 1) &= \frac{1 - P_{00}}{1 - P_{11} + 1 - P_{00}} \\
P(S_{t-1} = 0) &= \frac{1 - P_{11}}{1 - P_{11} + 1 - P_{00}};
\end{aligned}$$

take the derivatives in w and divide to get:

$$P(S_{t-1}|y_{t-1}=w)=\frac{P_{11}(1-P_{00})f_1(w)+(1-P_{00})(1-P_{11})f_0(w)}{(1-P_{00})f_1(w)+(1-P_{11})f_0(w)},$$

and the result follows.

Special cases include mean switching, with y_t i.i.d. conditionally on S and $\mu_{S_t}=S_t\mu_1+(1-S_t)\mu_2$, and stochastic volatility, with y_t i.i.d. conditionally on $Var(y_t)=\sigma^2_{S_t}$, and $\sigma^2_{S_t}=S_t\sigma^2_1+(1-S_t)\sigma^2_2$.

This example can easily be extended: as long as the distribution of y_t conditional on $S_t, y_{t-1}, \ldots, y_1$ has finite memory in y, so that there exists an m with

$$F(y_t|S_t, y_{t-1}, \ldots, y_1) \equiv F(y_t|S_t, y_{t-1}, \ldots, y_{t-m}),$$

then the variable $z_t=(y_{t-1}, \ldots, y_1)$ can be split into $x_t=(y_{t-1}, \ldots, y_{t-m})$ and $w_t=(y_{t-m-1}, \ldots, y_1)$. Thus Assumption 1 holds in any model in which the observed trajectory is a finite-order autoregressive conditionally on the hidden Markov chain. This is the case in most of the models in the literature.

2.3 Unobserved heterogeneity

Consider agents of unobserved type $s=0$ or $s=1$, and the following, fairly typical model with unobserved heterogeneity:

$$y=f(s,z,u).$$

Then Assumption 1 is implied by the following two restrictions:

1. for all (s,z,u),

$$f(s,z,u)=f(s,x,u);$$

   or equivalently,

$$y \perp\!\!\!\perp w \mid (s,x,u).$$

2. the distribution of u is conditionally independent of w:

$$u \perp\!\!\!\perp w \mid (x,s),$$

Assumption 2 then requires that w give some information on s, conditional on x. Many models used in the literature satisfy these assumptions. A random effects multinomial choice model (such as a mixed logit model) of consumer demand for instance would be

$$y=\arg\max_{k=1,\ldots,m}(v_k(z,s)+u_k),$$

with the u's are iid draws conditionally on z and s, so that item 2 above trivially holds. Item 1 holds if the mean utilities u_k do not depend on w, and Assumption 2 requires that the distribution of s conditional on z depend on w.

Thus what is crucial here, not surprisingly, is that there exist regressors w that do not enter mean utilities or random utility terms but that (loosely speaking) are correlated with the unobserved type s, as in (but not limited to) an auxiliary statistical model

$$s = 1 \text{ iff } \eta > P(z),$$

with η independent of w given x, and $P(z)$ depending on w.

In the consumer demand model for instance, this would hold if preferences can be well-approximated by a mixture of two types, whose proportion in each submarket (say) depends on geographical variables w, which do not enter their utility. Or, it applies if policies vary across subpopulations, but the policy variable does not appear in the utility function. In a labor supply model, the distribution of labor disutilities could vary across several groups w of observations. Note again that while such assumptions may be more or less convincing in a given application, they are testable.

Dynamic programming models often include "types": agents who are unobservably different, and whose differences are persistent over time. In a labor supply model for instance, the unobserved disutility of labor of a given agent can be approximated by breaking it into a time-invariant component (the type) and an iid component (or shock.) Then, just as in the regime switching model, past observed labor supplies give information on the type, while they are excluded from the distribution of labor supply given type.

We now turn to a less obvious example, inspired by the empirical industrial organization literature. Consider an oligopoly with N firms. Each firm i operates with constant marginal cost of production c_i and faces demand $D_i(p_i, p_{-i}, s)$, where the demand parameter s can take on two values $\overline{s} > \underline{s}$.

The timing of the game and the information structure are the following:

- Costs c_i are realized and observed by all firms.
- The firms simultaneously choose p_i to maximize their expected profits.
- Then s is realized.

- The econometrician later observes costs, prices, market shares, and profits of all firms:

$$\begin{cases} \tilde{c}_i &= c_i + \epsilon_{ci} \\ \tilde{p}_i &= p_i + \epsilon_{pi} \\ \tilde{D}_i &= D_i(p_i, p_{-i}, s) + \epsilon_{Di} \\ \tilde{\pi}_i &= (p_i - c_i) D_i(p_i, p_{-i}, s) + \epsilon_{\pi i}. \end{cases}$$

Let $\tilde{D}$, $\tilde{p}$, $\tilde{c}$, $\tilde{\pi}$ be the vectors of observed market shares, prices, costs, profits. Then, we have:

$$F(\tilde{D}|\tilde{p}, \tilde{\pi}, \tilde{c}) = F(\tilde{D}|\tilde{\pi}, \tilde{p}, \tilde{c}, s = \overline{s}) \Pr(s = \overline{s}|\tilde{\pi}, \tilde{p}, \tilde{c}) + F(\tilde{D}|\tilde{\pi}, \tilde{p}, \tilde{c}, s = \underline{s}) \Pr(s = \underline{s}|\tilde{\pi}, \tilde{p}, \tilde{c}).$$

Assumption 1 is satisfied under the following conditions:

1. Prices are observed by the econometrician without measurement error: $\epsilon_{pi} \equiv 0$.
2. The measurement error on market shares ϵ_{Di} and on costs ϵ_{ci} are independent of the measurement error on profits $\epsilon_{\pi i}$, conditional on $\tilde{c}_i$ and s.

When these conditions hold, we have the desired structure:

$$F(\tilde{D}|\tilde{p}, \tilde{\pi}, \tilde{c}) = F(\tilde{D}|\tilde{p}, \tilde{c}, s = \overline{s}) \Pr(s = \overline{s}|\tilde{\pi}, \tilde{p}, \tilde{c}) + F(\tilde{D}|\tilde{p}, \tilde{c}, s = \underline{s}) \Pr(s = \underline{s}|\tilde{\pi}, \tilde{p}, \tilde{c}),$$

where in the notation of the general structure above, the outcome y is observed demand $\tilde{D}$, the instrument w is observed profits $\tilde{\pi}$, and x consists of the vector $(\tilde{p}, \tilde{c})$ of observed prices and costs.

Note that condition 2 above can be realistically assumed if the information on profits comes from a later source, as in the case of Hendricks, Pinkse, and Porter (2003) where ex-post information is obtained on the value of oil tracts in wildcat lease contracts. Part 1 also is crucial: if prices were observed with error, then observing $\tilde{\pi}$ would give information on $\tilde{D}$, even after conditioning on observed prices and costs.

3 Partial Identification Results

We first exhaust the identifying power of assumption 1 before considering full nonparametric identification results in section 4. We present here partial identification results for the mixture weights and the mixture components, when only assumptions 1 and 2 hold.

3.1 Identifying the Weights

Now take any x in the set $\mathcal{X}$ of Definition 2. By construction, for such an x there exist two values of w which imply different cdfs of y conditional on x. Take any two such values and denote them $w_0(x)$ and $w_1(x)$. Clearly,

$$\lambda(w_0(x), x) \neq \lambda(w_1(x), x)$$

since otherwise the cdf of y conditional on $(w_k(x), x)$ would not depend on $k = 0, 1$ and x could not be in $\mathcal{X}$.

Under assumption 1, we can write for any y and any $z = (w, x)$:

$$\frac{F(y|z) - F(y|w_0(x), x)}{F(y|w_1(x), x) - F(y|w_0(x), x)} = \frac{\lambda(z) - \lambda(w_0(x), x)}{\lambda(w_1(x), x) - \lambda(w_0(x), x)}. \tag{3.1}$$

Hence, denoting

$$\Lambda(z) := \frac{F(y|z) - F(y|w_0(x), x)}{F(y|w_1(x), x) - F(y|w_0(x), x)},$$

any weight λ that rationalizes the data can only differ from Λ by an unknown pair $(\phi(x), \psi(x))$, which we called a (v, M) transform in section 1:

$$\lambda(z) = \phi(x) + \psi(x)\Lambda(z).$$

Note that $\phi(x)$ and $\psi(x)$ are related to the values of λ in $(w_k(x), x)$ according to the following very simple formulae: it follows from the definition of Λ that

$$\phi(x) = \lambda(w_0(x), x) \text{ and } \psi(x) = \lambda(w_1(x), x) - \lambda(w_0(x), x).$$

Thus $\phi(x) \geq 0$, but $\psi(x)$ may be negative. For the pair (ϕ, ψ), hence the (v, M) transform to be admissible, the corresponding λ needs to be a valid probability. It is easy to obtain identification regions for the functions ϕ and ψ. Denote $\overline{\Lambda}(x) = \sup_w \Lambda(w, x)$ and $\underline{\Lambda}(x) = \inf_w \Lambda(w, x)$. Then the following constraints for ϕ and ψ result from $0 \leq \lambda(z) \leq 1$:

$$\begin{aligned} 0 &\leq \phi(x) + \psi(x)\overline{\Lambda}(x) \leq 1 \\ 0 &\leq \phi(x) + \psi(x)\underline{\Lambda}(x) \leq 1. \end{aligned}$$

Equivalently,

$$\begin{aligned} -\psi(x)\overline{\Lambda}(x) \leq \quad & \phi(x) \quad \leq 1 - \psi(x)\overline{\Lambda}(x) \\ -\psi(x)\underline{\Lambda}(x) \leq \quad & \phi(x) \quad \leq 1 - \psi(x)\underline{\Lambda}(x). \end{aligned}$$

and finally

$$-\min(\psi(x)\overline{\Lambda}(x), \psi(x)\underline{\Lambda}(x)) \leq \phi(x) \leq 1 - \max(\psi(x)\overline{\Lambda}(x), \psi(x)\underline{\Lambda}(x)). \tag{3.2}$$

The inequalities in (3.2) completely define all admissible (v, M) transforms, and therefore they also define the partially identified regions for λ. Note that they immediately imply

$$\max(\psi(x)\overline{\Lambda}(x), \psi(x)\underline{\Lambda}(x)) - \min(\psi(x)\overline{\Lambda}(x), \psi(x)\underline{\Lambda}(x)) \leq 1$$

and therefore

$$|\psi(x)| \leq \frac{1}{\overline{\Lambda}(x) - \underline{\Lambda}(x)}.$$

This last inequality on $\psi(x)$ shows the impact of the variation of the cdf of y given z that is explained by w. If w strongly shifts the distribution of y given x , then it is clear from the definition of Λ that the bounds $\overline{\Lambda}(x)$ and $\underline{\Lambda}(x)$ will be further apart; then $\psi(x)$ will be constrained to a smaller interval, so that the variations of λ in w will be pinned down more closely.

Figure 1 represents the constraints on the pair $(\psi(x), \phi(x))$, suppressing the dependence in x for simplicity. It illustrates the point just made: a larger support for w will lead to an increase in $\overline{\Lambda} - \underline{\Lambda}$, and hence to a smaller identification region.

3.2 Identifying the Components

Once we settle on values for ϕ and ψ that satisfy (3.2) and thus on a λ that rationalizes the data, the mixture components obtain immediately. For all $x \in \mathcal{X}$,

$$\Delta(y|x) = F_1(y|x) - F_0(y|x) = \frac{F(y|w_1(x), x) - F(y|w_0(x), x)}{\lambda(w_1(x), x) - \lambda(w_0(x), x)}.$$

This also shows that $\Delta(y|x)$ is identified up to a multiplicative function of x. In fact, denoting $\tilde{\Delta}(y|x) = F(y|w_1(x), x) - F(y|w_0(x), x)$, we have

$$\Delta(y|x) = \frac{\tilde{\Delta}(y|x)}{\psi(x)},$$

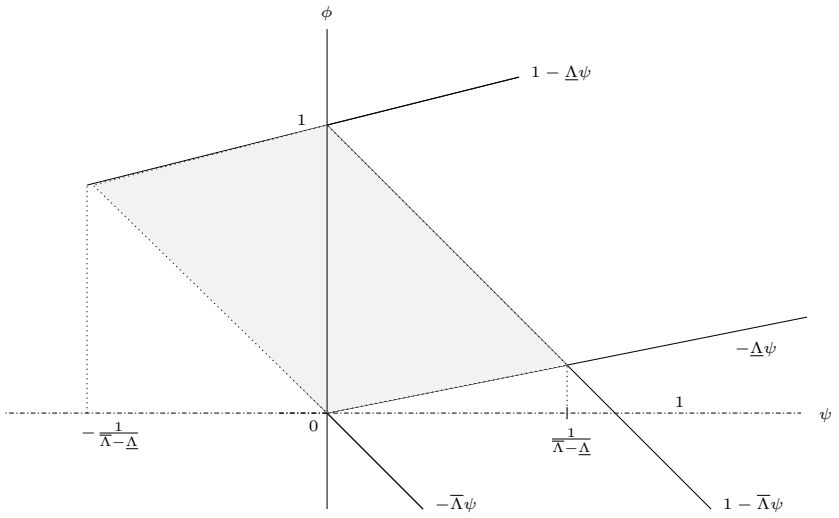

Figure 1: The shaded area is the identified region for the pair (ψ, ϕ) in a case where $\underline{\Lambda} < 0$ and $\overline{\Lambda} > 1$. In the binary regressor illustration with missing data, we shall see a case where $\underline{\Lambda} < 0$ and $\overline{\Lambda} = 1$. The only remaining case possible is $\underline{\Lambda} = 0$ and $\overline{\Lambda} = 1$, when the region is uninformative.

with the function ψ defined above.

The partial identification region for the mixture components can now be described as follows. By construction, we have

$$\begin{aligned} F_0(y|x) &= F(y|w_0(x), x) - \lambda(w_0(x), x)\Delta(y|x), \\ F_1(y|x) &= \Delta(y|x) + F_0(y|x) \\ &= F(y|w_0(x), x) + [1 - \lambda(w_0(x), x)]\Delta(y|x). \end{aligned}$$

By definition, $\phi(x) = \lambda(w_0(x), x)$ and $\psi(x) = \lambda(w_1(x), x) - \lambda_0(w_0(x), x)$. Hence the mixture components can be written

$$\begin{aligned} F_0(y|x) &= F(y|w_0(x), x) - \frac{\phi(x)}{\psi(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)], \\ F_1(y|x) &= F(y|w_0(x), x) + \frac{1 - \phi(x)}{\psi(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)]. \end{aligned}$$

and the identified region for the pair $(-\phi(x)/\psi(x), (1 - \phi(x))/\psi(x))$ is given in figure 2, where, as before, the dependence on x has been suppressed for simplicity.

A consequence of the identification of Δ up to scale is that some quantities of interest are in fact point identified without further assumption. For instance, take any function g of y such that $\mathbb{E}_{F_1} g - \mathbb{E}_{F_0} g \neq 0$. Then for any other function f of y, it is straightforward to

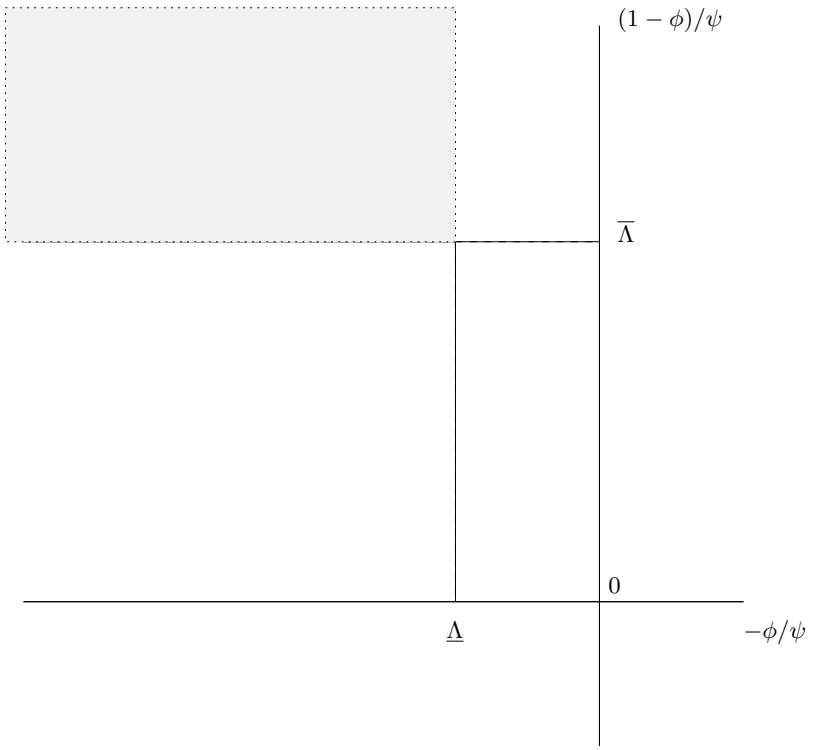

Figure 2: The shaded orthant is the identified region for the pair $(-\phi/\psi, (1-\phi)/\psi)$ which parameterize the two component distributions F_1 and F_0.

see that

$$r(x) := \frac{\mathbb{E}_{F_1} f - \mathbb{E}_{F_0} f}{\mathbb{E}_{F_1} g - \mathbb{E}_{F_0} g} \tag{3.3}$$
$$= \frac{\int f(y) dF(dy|w_1(x), x) - \int f(y) dF(dy|w_0(x), x)}{\int g(y) dF(dy|w_1(x), x) - \int g(y) dF(dy|w_0(x), x)}$$

One possible interpretation of this result is in term of "relative average effects of the true regressor": under Assumptions 1 and 2, we can compare the scale of the effects of the true regressor on various outcomes.

3.3 Illustration: Mismeasured Binary Regressor

The construction of ϕ and ψ requires that $x \in \mathcal{X}$, which implies the existence of two suitable values $w_0(x)$ and $w_1(x)$ for the instrument w. But if w can only take these two values, (3.2) in fact does not restrict λ in any way: Λ only takes the values 0 and 1, so that (3.2) boils down to

$$-\min(0, \psi(x)) \leq \phi(x) \leq 1 - \max(0, \psi(x)),$$

which defines a triangle in (ψ, ϕ) space with corners $(-1, 1), (0, 1)$, and $(1, 0)$. Easy calculations show that this maps into the $[0, 1] \times [0, 1]$ square in $(\lambda(w_0(x), x), \lambda(w_1(x), x))$ space, so that the mixture weights could be anything. In this case partial identification does not achieve much: we need more values for w. Note, however, that (3.3) still holds and so it is possible to make some statements about ratios of differences.

This can be illustrated in the regression model with binary regressor. First consider the case where all individuals are classified, with possible misclassification error, so that reported status T only takes values $T = 0$ and $T = 1$. Then (dropping $x \in \mathcal{X}$ from the notation)

$$\Lambda(T) = \frac{F(y|T) - F(y|T = 0)}{F(y|T = 1) - F(y|T = 0)}$$

(which, again, does not depend on y given Assumption 1) can only take the values zero or one. As explained above, this is entirely uninformative about the mixture probabilities $\lambda(T) = \Pr(T^* = 1|T)$. The data does not tell us anything about the mismeasurement process. Nor can we deduce much about the components F_1 and F_0; we cannot go beyond quantities like the ratio (3.3) above.

Now consider the case where some classification information is missing, so that reported T can take three distinct values: $T = 1$, $T = 0$ and $T = \varnothing$ if the individual is not classified.

Then we can define (for instance)

$$\Lambda(T) = \frac{F(y|T) - F(y|T = \varnothing)}{F(y|T = 1) - F(y|T = \varnothing)}$$

which can take three distinct values: $\Lambda(1) = 1$, $\Lambda(\varnothing) = 0$ and $\Lambda(0)$.

Suppose $\Lambda(0) < 0$, so that the maximum and minimum values in our partial identification analysis are now $\overline{\Lambda} = \Lambda(1) = 1$ and $\underline{\Lambda} = \Lambda(0) < 0$.

Now the restrictions on ϕ and ψ are

$$-\min(\psi, \psi\Lambda(0)) \leq \phi \leq 1 - \max(\psi, \psi\Lambda(0));$$

and the partial identification regions are no longer trivial. Denote $L = 1/(1 - \Lambda(0))$ so that $0 < L < 1$; then (ψ, ϕ) must be in the trapeze defined by the corners $(-L, L), (0, 0), (0, 1)$ and $(L, 1 - L)$.

The mixture components are given for any y by

$$F(y|T^* = 1) = F(y|T = \varnothing) + \frac{1 - \phi}{\psi}(F(y|T = 1) - F(y|T = \varnothing))$$

and

$$F(y|T^* = 0) = F(y|T = \varnothing) - \frac{\phi}{\psi}(F(y|T = 1) - F(y|T = \varnothing)).$$

This example also shows how easy it is to test for our exclusion restriction: one could compute

$$\Lambda(0) = \frac{F(y|T = 0) - F(y|T = \varnothing)}{F(y|T = 1) - F(y|T = \varnothing)}$$

for several values of y for instance and check that they give similar answers, as they should.

4 Point Identification

We now turn to full nonparametric identification of mixture weights and mixture components based on tail conditions that we illustrate and interpret in our main examples.

Definition 3 *Call $\mathcal{V}_-$ the set of values of x such that $F_1(y|x)/F_0(y|x) \to_{y \to -\infty} 0$, $\mathcal{V}_+$ the set of values of x such that $(1 - F_0(y|x))/(1 - F_1(y|x)) \to_{y \to +\infty} 0$.*

With this definition, we have

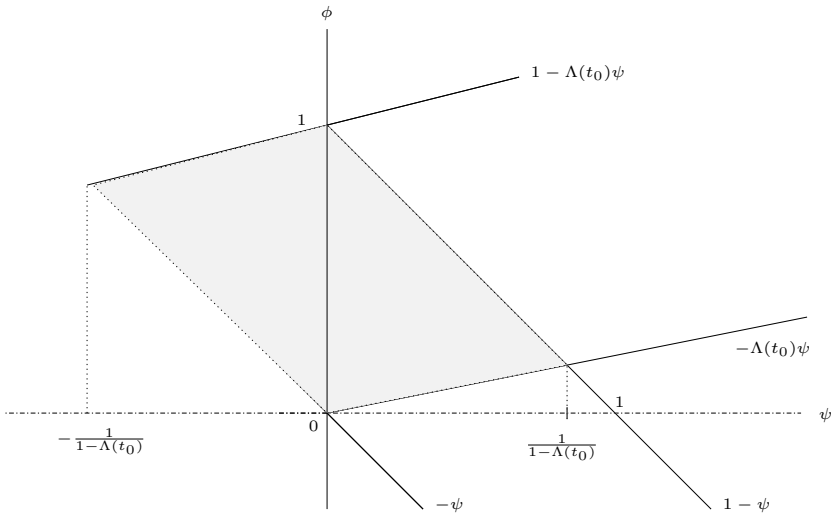

Figure 3: Identified region for component distributions with mismeasured binary regressor

Lemma 3 *Under assumption 1, for all $x \in \mathcal{X} \cap \mathcal{V}_- \cap \mathcal{V}_+$, λ, F_1 and F_0 are nonparametrically identified.*

Proof Let $w_0(x)$ and $w_1(x)$ be chosen as in section 3. Since $x \in \mathcal{X}$, $\lambda(w_0(x), x)$ can be chosen different from zero or one. Assumption 1 implies that on $\mathcal{X}$, we have for all y such that $F_0(y|x) \neq 0$ (which will hold for y large enough if $x \in \mathcal{V}_- \cap \mathcal{V}_+$)

$$F(y|z) = F_0(y|x)\left(1 + \lambda(z)\left(\frac{F_1(y|x)}{F_0(y|x)} - 1\right)\right)$$

so that

$$\zeta(y, x) := \frac{F(y|w_1(x), x)}{F(y|w_0(x), x)} = \frac{1 + \lambda(w_1(x), x)\left(\frac{F_1(y|x)}{F_0(y|x)} - 1\right)}{1 + \lambda(w_0(x), x)\left(\frac{F_1(y|x)}{F_0(y|x)} - 1\right)}.$$

Hence, calling $\zeta_-(x) = \lim_{y \to -\infty} \zeta(y, x)$, we have

$$\frac{1 - \lambda(w_1(x), x)}{1 - \lambda(w_0(x), x)} = \zeta_-(x). \tag{4.1}$$

Similarly,

$$1 - F(y|z) = (1 - F_0(y|x))\left(1 - \lambda(z) + \lambda(z)\left(\frac{1 - F_1(y|x)}{1 - F_0(y|x)}\right)\right)$$

so that

$$\xi(y,x) := \frac{1-F(y|w_1(x),x)}{1-F(y|w_0(x),x)} = \frac{1-\lambda(w_1(x),x)+\lambda(w_1(x),x)\left(\frac{1-F_1(y|x)}{1-F_0(y|x)}\right)}{1-\lambda(w_0(x),x)+\lambda(w_0(x),x)\left(\frac{1-F_1(y|x)}{1-F_0(y|x)}\right)}.$$

Hence, calling $\xi_+(x) = \lim_{y\to+\infty}\xi(y,x)$, on $\mathcal{X}\cap\mathcal{V}_-\cap\mathcal{V}_+$ we have

$$\frac{\lambda(w_1(x),x)}{\lambda(w_0(x),x)} = \xi_+(x) \tag{4.2}$$

Combining equations (4.1) and (4.2) and noting that $\zeta_-(x) \neq \xi_+(x)$ (otherwise we would have $\lambda(w_0(x),x) = \lambda(w_1(x),x)$, contradicting $x\in\mathcal{X}$) gives the result:

$$\begin{aligned}\phi(x) &= \lambda(w_0(x),x) = \frac{1-\zeta_-(x)}{\xi_+(x)-\zeta_-(x)}\\ \psi(x) &= \lambda(w_1(x),x)-\lambda(w_0(x),x) = \frac{(1-\xi_+(x))(1-\zeta_-(x))}{\zeta_-(x)-\xi_+(x)}.\end{aligned}$$

To illustrate this result, first go back to the mismeasured binary regressor model. Then the assumption needed for point identification of component distributions and mixture weights is the following:

$$\frac{F(y|T^*=1)}{F(y|T^*=0)}\to 0 \text{ as } y\to-\infty \quad\text{and}\quad \frac{1-F(y|T^*=0)}{1-F(y|T^*=1)}\to 0 \text{ as } y\to+\infty.$$

In other words, the conditional distribution of outcomes given $T^*=1$ dominates the right tail and that with $T^*=0$ dominates the left tail. Note that nothing is required of the rest of the distribution Of course, the roles of right and left tail could be reversed if the model under study makes it more natural.

If in particular the mismeasurement is such that $y \perp\!\!\!\perp T|T^*$, outcomes are normally distributed conditionally on the regressor, and the regressor only shifts the mean outcome, then this condition is satisfied; in fact in that case $F(y|T^*=1)$ and $F(y|T^*=0)$ are normal with identical variances and different means, which implies the tail conditions we need, as shown in the following simple lemma (which we include here for completeness, as we could not find a simple reference in the literature).

Lemma 4 *Let Φ be the cumulative distribution of the standard normal random variable, and $\mu_0<\mu_1$ two real numbers. Then*

$$\frac{1-\Phi\left(\frac{y-\mu_0}{\sigma}\right)}{1-\Phi\left(\frac{y-\mu_1}{\sigma}\right)}\to 0 \text{ as } y\to+\infty \text{ and } \frac{\Phi\left(\frac{y-\mu_1}{\sigma}\right)}{\Phi\left(\frac{y-\mu_0}{\sigma}\right)}\to 0 \text{ as } y\to-\infty.$$

Proof By L'Hôpital's rule, the result follows if the densities have the required limiting ratios, which is verified as follows. Let ϕ be the density of a standard normal random variable. Then, with $K = \exp((\mu_1^2 - \mu_0^2)/(2\sigma^2))$, we have $\phi\left(\frac{y-\mu_0}{\sigma}\right)/\phi\left(\frac{y-\mu_1}{\sigma}\right) = K\exp(2y(\mu_1-\mu_0)/(2\sigma^2))$, which tends to 0 as y tends to $+\infty$. The other case is treated identically.

The previous lemma shows that the tail dominance of definition 3 necessary for nonparametric identification of the model is satisfied in the case of normally distributed outcomes, where the regressor only affects location. This extends to more general location models, as shown below.

Lemma 5 *In the location model $F(y|T^*, x) = F(y - m(T^*)|x)$, where m is a decreasing function and $-\ln(1 - F(y|x))$ (resp. $-\ln F(y|x)$) has a derivative that grows unboundedly for y large enough (resp. small enough), we have $x \in \mathcal{X} \cap \mathcal{V}_- \cap \mathcal{V}_+$ so that the model is nonparametrically identified.*

Proof of lemma 5: Let $f(y|T^* = j, x) = f(y - m_j, x)$, $j = 0, 1$ be the density of the two regimes, with $m_1 > m_0$. Let $g(x, y) = -\ln(1 - F(y|x))$. The first derivative of g is increasing to $+\infty$ by assumption. Hence

$$\frac{1 - F(y - m_0|x)}{1 - F(y - m_1|x)} = \exp[g(x, y - m_1) - g(x, y - m_0)]$$

tends to 0 when $y \to +\infty$.

Under the identifying assumption, we can write

$$\zeta_- = \lim_{y\to-\infty} F(y|T = 1)/F(y|T = 0) = \Pr(T^* = 0|T = 1)/\Pr(T^* = 0|T = 0)$$

and

$$\xi_+ = \lim_{y\to+\infty} [1 - F(y|T = 1)]/[1 - F(y|T = 0)] = \Pr(T^* = 1|T = 1)/\Pr(T^* = 1|T = 0),$$

from which the mixture weights and the component distributions are identified.

The tail conditions also hold in most variants of the regime switching model which have been considered in the literature. Take an application to macroeconomic data for instance. Then $\mathcal{V}_- \cap \mathcal{V}_+$ corresponds to the set of conditioning points where the regime associated with F_1 dominates the upper tail, hence is the *expansionary* regime, and the regime associated with F_0 dominates the lower tail, hence the *contraction* regime. Note that this would be

the case, for instance, with F_1 and F_0 normal with identical variances and different means (as in the original model of Hamilton (1989)).

In some applications it may be too much to ask for tail conditions at both ends. If a tail condition holds only in one tail, then we are back to partial identification, but more is point identified than in section 3. More precisely, focus for instance on tail dominance in the right tail. On the larger set $\mathcal{X} \cap \mathcal{V}_+$ of conditioning points, the following lemma shows that the dominated regime is fully identified.

Lemma 6 *Under assumption 1, for all $x \in \mathcal{X} \cap \mathcal{V}_+$, F_0 is point identified, whereas F_1 and λ are identified up to a constant.*

Proof As above, we have

$$\frac{\lambda(w_1(x), x)}{\lambda(w_0(x), x)} = \xi_+(x).$$

Since $\lambda(z) = \phi(x) + \psi(x)\Lambda(z)$, we have

$$\frac{\phi(x) + \psi(x)\Lambda(w_1(x), x)}{\phi(x) + \psi(x)\Lambda(w_0(x), x)} = \xi_+(x),$$

from which it follows that

$$\phi(x) = \tilde{\kappa}(x)\psi(x),$$

with (remember that $\xi_+(x) \neq 1$ for $x \in \mathcal{X}$)

$$\tilde{\kappa}(x) = \frac{\Lambda(w_1(x), x) - \xi_+\Lambda(w_0(x), x)}{\xi_+ - 1}$$

Hence, we have

$$\begin{aligned} F(y|z) &= \lambda(z)\Delta(y|x) + F_0(y|x) \\ &= (\phi(x) + \psi(x)\Lambda(z))\frac{\tilde{\Delta}(y|x)}{\psi(x)} + F_0(y|x) \\ &= \psi(x)(\tilde{\kappa}(x) + \Lambda(z))\frac{\tilde{\Delta}(y|x)}{\psi(x)} + F_0(y|x) \\ &= (\tilde{\kappa}(x) + \Lambda(z))\tilde{\Delta}(y|x) + F_0(y|x) \end{aligned}$$

and F_0 is point identified.

The stochastic volatility model illustrates the usefulness of lemma 6. By definition, the regime with more volatility dominates in both tails; then we can resort to this result to prove that the regime with lower volatility is point identified.

5 Estimation

We now propose an estimator for the mixture components and mixture weights. This is based on the identifiable quantities

$$\zeta_-(x) := \lim_{y \to -\infty} \frac{F(y|w_1(x), x)}{F(y|w_0(x), x)}$$

and

$$\xi_+(x) := \lim_{y \to +\infty} \frac{1 - F(y|w_1(x), x)}{1 - F(y|w_0(x), x)}.$$

We already showed that

$$\begin{aligned} F_0(y|x) &= F(y|w_0(x), x) - \frac{\phi(x)}{\psi(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)], \\ F_1(y|x) &= F(y|w_0(x), x) + \frac{1 - \phi(x)}{\psi(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)], \end{aligned}$$

which, under the identifying assumptions of lemma 3 can be rewritten as

$$\begin{aligned} F_0(y|x) &= F(y|w_0(x), x) + \frac{1}{1 - \xi_+(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)], \\ F_1(y|x) &= F(y|w_0(x), x) + \frac{1}{1 - \zeta_-(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)]. \end{aligned}$$

For some diverging sequences $R \to \infty$ and $L \to -\infty$, we propose the following estimators for $\zeta_-(x)$ and $\xi_+(x)$.

Definition 4 (Estimators) *Define* $\hat{\zeta}_L(x) = \hat{F}(L|w_1(x), x)/\hat{F}(L|w_0(x), x)$ *and* $\hat{\xi}_R(x) = [1 - \hat{F}(R|w_1(x), x)]/[1 - \hat{F}(R|w_0(x), x)]$*, where* $\hat{F}$ *is a nonparametric estimator of the conditional cumulative distribution of* y *given* x *and* w*.*

We now give the statistical properties of our estimation strategy in the case of the mismeasured binary regressor, where the outcome distributions of interest are characterized by $F(y|T^* = 1, x)$ and $F(y|T^* = 0, x)$. An iid sample of individuals is available with their reported regressor values and their outcomes. For simplicity, we assume that T^* is the only observed regressor, so that x drops from the notation. The results extend trivially to the case with any number of additional discrete-valued regressors x.

Assumption 3 $((y_1, T_1), \ldots, (y_n, T_n))$ *is an iid sample.*

Remark 1 *The results can be extended to weakly dependent sequences using convergence results for tail empirical processes in Rootzén (2009), and to the case with conditioning information using more general local empirical process results in Einmahl and Mason (1997).*

For diverging sequences $R \to \infty$ and $L \to -\infty$, we have $\hat{\zeta}_L = \hat{F}(L|T=1)/\hat{F}(L|T=0)$ and $\hat{\xi}_R = [1-\hat{F}(R|T=1)]/[1-\hat{F}(R|T=0)]$. In the above expressions, $\hat{F}(y|T=1)$ and $\hat{F}(y|T=0)$ are the empirical cumulative distributions of the sample of the $T=1$ and $T=0$ categories respectively. For instance, $\hat{F}(y|T=1)$ is the fraction of outcomes of $T=1$ individuals in the sample that fall below y.

Definition 5 *The empirical distributions are defined as follows.* $\hat{F}(y|T=j) = \#\{1 \leq i \leq n : T_i = j, Y_i \leq y\}/n_j$ *where* $n_j = \#\{1 \leq i \leq n : T_i = j\}$, $j = 1, 0$. *The corresponding empirical processes are defined as* $\mathbb{G}_{n_j}(y|T=j) = \sqrt{n_j}(\hat{F}(y|T=j) - F(y|T=j))$, $j = 1, 0$.

Since $(1-\hat{\zeta}_L)^{-1} = -\hat{F}(L|T=0)/[\hat{F}(L|T=1) - \hat{F}(L|T=0)]$ and $(1-\hat{\xi}_R)^{-1} = [1-\hat{F}(R|T=0)]/[\hat{F}(R|T=1) - \hat{F}(R|T=0)]$, the resulting estimators for the outcome distributions are the following.

Definition 6 (Nonparametric estimators for the component distributions:)

$$\begin{aligned}
\hat{F}(y|T^*=0) &= \hat{F}(y|T=0) + [1-\hat{F}(R|T=0)]\frac{\hat{F}(y|T=1) - \hat{F}(y|T=0)}{\hat{F}(R|T=1) - \hat{F}(R|T=0)} \\
\hat{F}(y|T^*=1) &= \hat{F}(y|T=0) - \hat{F}(L|T=0)\frac{\hat{F}(y|T=1) - \hat{F}(y|T=0)}{\hat{F}(L|T=1) - \hat{F}(L|T=0)}.
\end{aligned}$$

As we have shown in section 4, the mixture weights are identified as $\lambda(T=1) = \xi_+(1-\zeta_-)/(\xi_+ - \zeta_-)$ and $\lambda(T=0) = (1-\zeta_-)/(\xi_+ - \zeta_-)$, we estimate them in the following way:

Definition 7 (Nonparametric estimators of mixture weights)

$$\begin{pmatrix} \hat{\lambda}(T=1) \\ \hat{\lambda}(T=0) \end{pmatrix} = g \begin{pmatrix} \hat{\xi}_R \\ \hat{\zeta}_L \end{pmatrix} \quad \textit{with} \quad g : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \frac{x(1-y)}{x-y} \\ \frac{1-y}{x-y} \end{pmatrix}$$

Notice that for these estimators to make sense, we need some outcomes of the units with $T=0$ to fall on the right of R and on the left of L. For statistical purposes, we need these numbers of outcomes to grow with n. Hence the following assumption on the sequences R and L:

Assumption 4 (Order statistics) *R (resp. L) is chosen as the statistic of order $r_n + 1$ (resp. $n_0 - l_n$), i.e. the $r_n + 1$-th largest (resp. l_n-th smallest) outcome value of the sample of individuals with $T = 0$, with $r_n/n_0 \to 0$ and $r_n/\sqrt{n_0 \ln\ln n_0} \to \infty$ (and similar conditions hold for l_n).*

Remark 2 *In practice, one would definitely want to choose R and L after examining the distributions of outcomes for observations with $T = 0$ and observations for $T = 1$. In particular, to prevent the estimator from taking infinite values, we should choose R and L such that $1 - \hat{F}(R|T = 1) \neq 1 - \hat{F}(R|T = 0)$ and $\hat{F}(L|T = 1) \neq \hat{F}(L|T = 0)$.*

We assume that reported regressor value is informative, but not perfectly correlated with actual value (in which case, the identification issue would disappear).

Assumption 5 *We have $1/2 < \Pr(T^* = 1|T = 1) < 1$ and $1/2 < \Pr(T^* = 0|T = 0) < 1$.*

Our identification strategy relies on tail dominance of the outcome distribution under $T = 1$ in the right tail and dominance of the outcome distribution under $T = 0$ in the left tail of the distribution[4]. For asymptotic normality of the proposed estimator, we need to assume this dominance holds at a certain rate.

Assumption 6 *$[1 - F(R|T^* = 0)]/[1 - F(R|T^* = 1)]) = o_p(1/\sqrt{r_n})$, and $F(L|T^* = 1)/F(L|T^* = 0) = o_p(1/\sqrt{l_n})$.*

Assumption 6 is relatively easy to check. It holds in particular in the following cases:

- The case where $F(y|T^* = 1)$ and $F(y|T^* = 0)$ have log concave tails (which includes, but is not restricted to Gaussian tails) as shown in lemma 7.
- The case where $F(y|T^* = 1)$ and $F(y|T^* = 0)$ have fat tails satisfying the conditions of lemma 8.

Lemma 7 (Case of log concave tails:) *If we have*

$$\begin{aligned} -\ln(1 - F(y|T^*)) &\sim \left(\frac{y}{\sigma_+(T^*)}\right)^{\alpha_+(T^*)} \quad as \quad y \to +\infty \\ -\ln F(y|T^*) &\sim \left(\frac{y}{\sigma_-(T^*)}\right)^{\alpha_-(T^*)} \quad as \quad y \to -\infty, \end{aligned}$$

with $\sigma_{-,+}(T^*) > 0$ and $\alpha_{-,+}(T^*) > 1$, then

- *$\alpha_+(T^* = 1) < \alpha_+(T^* = 0)$ or $[\alpha_+(T^* = 1) = \alpha_+(T^* = 0)$ and $\sigma_+(T^* = 1) > \sigma_+(T^* = 0)$],*

[4] Again, this could be reversed if the model under study calls for it.

- $\alpha_-(T^*=1) > \alpha_-(T^*=0)$ *or* $[\alpha_-(T^*=1) = \alpha_-(T^*=0)$ *and* $\sigma_-(T^*=1) < \sigma_-(T^*=0)]$,

jointly imply assumption 6.

Proof of lemma 7: In what follows, K is a generic constant and $\sim_p$ denotes first order equivalence in probability as $n \to \infty$.

We have $1 - F(R|T=0) = \Pr(T^*=1|T=0)[1-F(R|T^*=1)] + \Pr(T^*=0|T=0)[1-F(R|T^*=0)] = (1-\Pr(T^*=0|T=0))[1-F(R|T^*=1)](1+o_p(1))$ under assumption 5 and the tail conditions.

Moreover, by assumption 4, $r_n/n_0 = 1 - \hat{F}(R|T=0) = 1 - F(R|T=0) + \mathbb{G}_{n_0}(R|T=0)/\sqrt{n_0} = [1-F(R|T=0)] + O_{\text{a.s.}}(\sqrt{\ln\ln n_0/n_0})$ by the law of iterated logarithm. With the result of the previous paragraph, this yields $r_n/n_0 = K[1-F(R|T^*=1)](1+o_p(1))$.

Given the assumption on the tails of F, this yields $R \sim_p K(\ln n_0)^{1/\alpha(1)}$. Hence,

$$\frac{1-F(R|T^*=0)}{1-F(R|T^*=0)} \quad \sim_p \quad \exp\left(\left(\frac{R}{\sigma_+(1)}\right)^{\alpha_+(1)} - \left(\frac{R}{\sigma_+(0)}\right)^{\alpha_+(0)}\right).$$

The latter is of order $\exp\left\{-K(\ln n_0)^{\alpha_+(0)/\alpha_+(1)}\right\}$ when $\alpha_+(0) > \alpha_+(1)$, and of order $\exp\{-K\ln n_0\}$ when $\alpha_+(0) = \alpha_+(1)$ and $\sigma_+(1) > \sigma_+(0)$.

Lemma 8 (Case of Pareto tails) *Suppose* $\alpha_{0R} > \alpha_{1R} > 0$ *and* $\alpha_{1L} > \alpha_{0L} > 0$. *Denote by* $c > 0$ *a generic positive finite constant. Suppose the distribution of outcomes conditional on* $T^* = 1$ *has right (resp. left) tail index* α_{1R} *(resp.* α_{1L}*), namely* $1-F(y|T^*=1) \sim cy^{-\alpha_{1R}}$ *when* y *tends to* ∞ *(resp.* $F(y|T^*=1) \sim c(-y)^{-\alpha_{1L}}$ *as* y *tends to* $-\infty$*). Suppose similarly that the distribution of outcomes conditional on* $T^*=0$ *has right (resp. left) tail index* α_{0R} *(resp.* α_{0L}*), namely* $1-F(y|T^*=0) \sim cy^{-\alpha_{0R}}$ *when* y *tends to* ∞ *(resp.* $F(y|T^*=0) \sim c(-y)^{-\alpha_{0L}}$ *as* y *tends to* $-\infty$*). Then, under assumptions 4 and 5, assumption 6 holds when*

$$r_n = o\left(n_0^{\frac{2(\alpha_{0R}-\alpha_{1R})}{2\alpha_{0R}-\alpha_{1R}}}\right) \qquad \text{and} \qquad l_n = o\left(n_0^{\frac{2(\alpha_{1L}-\alpha_{0L})}{2\alpha_{1L}-\alpha_{0L}}}\right).$$

In particular, when $\alpha_{0R} = 2\alpha_{1R}$*, assumption 6 will be satisfied if* $r_n = o(n_0^{2/3})$*, and when* $\alpha_{0R} = 3\alpha_{1R}$*, assumption 6 will be satisfied if* $r_n = o(n_0^{4/5})$*, and similarly for* l_n.

Proof of lemma 8: First note that $1 - F(R|T=0) = \Pr(T^*=1|T=0)[1-F(R|T^*=1)] + \Pr(T^*=0|T=0)[1-F(R|T^*=0)] = (1-\Pr(T^*=0|T=0))[1-F(R|T^*=1)](1+o_p(1))$ under assumption 5 and $\alpha_{1R} < \alpha_{0R}$.

Now, by assumption 4, $r_n/n_0 = 1 - \hat{F}(R|T=0) = 1 - F(R|T=0) + \mathbb{G}_{n_0}(R|T=0)/\sqrt{n_0} = [1 - F(R|T=0)] + O_{\text{a.s.}}(\sqrt{\ln\ln n_0/n_0})$ by the law of iterated logarithm. With the result of the previous paragraph, this yields $r_n/n_0 = c[1 - F(R|T^*=1)](1 + o_p(1))$.

Finally, by the assumption of the lemma, $1 - F(R|T^*=1) \sim cR^{-\alpha_{1R}}$, hence $r_n/n_0 = cR^{-\alpha_{1R}}(1 + o_p(1))$.

The tail dominance requirement of assumption 6 is $[1 - F(R|T^*=0)]/[1 - F(R|T^*=1)] = o(r_n^{-1/2})$, which is therefore equivalent to $R^{\alpha_{1R}-\alpha_{0R}} = o(r_n^{-1/2})$ or

$$[(n_0/r_n)^{1/\alpha_{1R}}]^{\alpha_{1R}-\alpha_{0R}} = o(r_n^{-1/2}),$$

and the result follows. The case of the left tail is treated identically.

Finally, for the asymptotic treatment of the tail empirical process, we assume that the conditional cumulative outcome distribution functions given $T^* = 0, 1$ are invertible.

Assumption 7 *Both $y \to F(y|T^*=0)$ and $y \to F(y|T^*=1)$ are continuous and strictly increasing.*

Remark 3 *Notice that assumption 7 is very mild. It does not require the existence of moments for the outcome distributions.*

Under the previous assumptions, we have the following theorem.

Theorem 2 *Under assumptions 1-7, the centered and re-scaled estimator of definition 6*

$$\sqrt{r_n}\begin{pmatrix} \hat{F}(y|T^*=1) - F(y|T^*=1) \\ \hat{F}(y|T^*=0) - F(y|T^*=0) \end{pmatrix}$$

is asymptotically normal with mean zero and variance $[F(y|T=1) - F(y|T=0)]^2 A^{-1} V$ with

$$V = \begin{pmatrix} \xi_+^2 + \rho\xi_+ & 0 \\ 0 & \zeta_-^2 + \rho\zeta_- \end{pmatrix} \quad \textit{and} \quad A = \begin{pmatrix} (1-\xi_+)^2 & 0 \\ 0 & (1-\zeta_-)^2 \end{pmatrix}$$

where $\rho = Pr(T=0)/Pr(T=1)$.

The corresponding rescaled and centered mixture weights estimator of definition 7

$$\sqrt{r_n}\begin{pmatrix} \hat{\lambda}(T=1) - \lambda(T=1) \\ \hat{\lambda}(T=0) - \lambda(T=0) \end{pmatrix}$$

is asymptotically normal with mean zero and variance BVB^t with

$$B = \frac{1}{(\xi_+ - \zeta_-)^2} \begin{pmatrix} \xi_+(1-\xi_+) & (1-\xi_+)(1-\zeta_-) \\ 1-\xi_+ & \zeta_- - 1 \end{pmatrix}$$

Proof of Theorem 2 In all that follows, the stochastic dominance relations are uniform with respect to y. Consider first $\hat{F}(y|T^*=0) - F(y|T^*=0)$. We use the following notation.

$$\begin{aligned} \hat{F}(y|T^*=0) &= \hat{F}(y|T=0) + \hat{K}_n \hat{D} \\ F(y|T^*=0) &= F(y|T=0) + KD, \end{aligned}$$

where

$$\begin{aligned} D &= F(y|T=1) - F(y|T=0) \\ \hat{D} &= \hat{F}(y|T=1) - \hat{F}(y|T=0) \end{aligned}$$

and

$$\begin{aligned} K &= \frac{1}{1-\xi_+} \\ \hat{K}_n &= \frac{1}{1-\hat{\xi}_R}, \end{aligned}$$

where $\xi_R = S_1/S_0$ and $\hat{\xi}_R = \hat{S}_1/\hat{S}_0$ and where S_j denotes the survival functions for $j = 0, 1$:

$$\begin{aligned} \hat{S}_j &= 1 - \hat{F}(R|T=j) \\ S_j &= 1 - F(R|T=j). \end{aligned}$$

Finally, we use standard notation for the empirical process, namely:

$$\begin{aligned} \mathbb{G}_{n_0}(y) &= \sqrt{n_0}\left(\hat{F}(y|T=0) - F(y|T=0)\right) \\ \mathbb{G}_{n_1}(y) &= \sqrt{n_1}\left(\hat{F}(y|T=1) - F(y|T=1)\right) \end{aligned}$$

We these notations, by the Glivenko-Cantelli Theorem, we have

$$\begin{aligned} & \sqrt{r_n}\left[\hat{F}(y|T^*=0) - F(y|T^*=0)\right] \\ = \; & \sqrt{r_n}\left(\hat{F}(y|T=0) - F(y|T=0)\right) + \hat{D}\sqrt{r_n}(\hat{K}_n - K) + K\sqrt{r_n}(\hat{D} - D) \\ = \; & \hat{D}\sqrt{n}(\hat{K}_n - K) + o_p(1). \end{aligned}$$

Now

$$\begin{aligned}\xi_R &= \frac{\Pr(T^*=1|T=1)[1-F(R|T^*=1)]+(1-\Pr(T^*=1|T=1))[1-F(R|T^*=0)]}{(1-\Pr(T^*=0|T=0))[1-F(R|T^*=1)]+\Pr(T^*=0|T=0)[1-F(R|T^*=0)]},\\ &= \frac{\Pr(T^*=1|T=1)+[1-\Pr(T^*=1|T=1)]\frac{1-F(R|T^*=0)}{1-F(R|T^*=1)}}{[1-\Pr(T^*=0|T=0)]+\Pr(T^*=0|T=0)\frac{1-F(R|T^*=0)}{1-F(R|T^*=1)}}.\end{aligned}$$

Hence $\xi_R - \xi_+$ is equal to

$$\begin{aligned}&\xi_R - \frac{\Pr(T^*=1|T=1)}{1-\Pr(T^*=0|T=0)}\\ =\ & \frac{[1-\Pr(T^*=0|T=0)]\left(\Pr(T^*=1|T=1)+[1-\Pr(T^*=1|T=1)]\frac{1-F(R|T^*=0)}{1-F(R|T^*=1)}\right)}{[1-\Pr(T^*=0|T=0)]\left([1-\Pr(T^*=0|T=0)]+\Pr(T^*=0|T=0)\frac{1-F(R|T^*=0)}{1-F(R|T^*=1)}\right)}\\ -\ & \frac{\Pr(T^*=1|T=1)\left([1-\Pr(T^*=0|T=0)]+\Pr(T^*=0|T=0)\frac{1-F(R|T^*=0)}{1-F(R|T^*=1)}\right)}{[1-\Pr(T^*=0|T=0)]\left([1-\Pr(T^*=0|T=0)]+\Pr(T^*=0|T=0)\frac{1-F(R|T^*=0)}{1-F(R|T^*=1)}\right)}\\ =\ & \frac{[1-\Pr(T^*=0|T=0)-\Pr(T^*=1|T=1)]\frac{1-F(R|T^*=0)}{1-F(R|T^*=1)}}{[1-\Pr(T^*=0|T=0)]\left([1-\Pr(T^*=0|T=0)]+\Pr(T^*=0|T=0)\frac{1-F(R|T^*=0)}{1-F(R|T^*=1)}\right)},\end{aligned}$$

which is $O_p([1-F(R|T^*=0)]/[1-F(R|T^*=1)])$ because $0 < \Pr(T^*=0|T=0) < 1$ under assumption 5. Hence we have shown that $\sqrt{r_n}(\xi_R - \xi_+)$ is $O_p(\sqrt{r_n}[1-F(R|T^*=0)]/[1-F(R|T^*=1)])$, which is $o_p(1)$ by assumption 6.

We can therefore write

$$\begin{aligned}\sqrt{r_n}(\hat{\xi}_R - \xi_+) &= \sqrt{r_n}(\hat{\xi}_R - \xi_R + \xi_R - \xi_+)\\ &= \sqrt{r_n}(\hat{\xi}_R - \xi_R) + o_p(1).\end{aligned}$$

But

$$\begin{aligned}\sqrt{r_n}(\hat{\xi}_R - \xi_R) &= \sqrt{r_n}\left(\frac{\hat{S}_1}{\hat{S}_0} - \frac{S_1}{S_0}\right)\\ &= \sqrt{r_n}\frac{1}{\hat{S}_0}\left[(\hat{S}_1 - S_1) - \xi_R(\hat{S}_0 - S_0)\right]\\ &= \sqrt{\frac{r_n}{n_0}}\frac{1}{\hat{S}_0}\left[\xi_R\mathbb{G}_{n_0}(R) - \mathbb{G}_{n_1}(R)\right]\end{aligned}$$

By construction, $\hat{S}_0 = r_n/n_0$, so that

$$\sqrt{r_n}(\hat{\xi}_R - \xi_R) = \sqrt{\frac{n_0}{r_n}}\left[\xi_R\mathbb{G}_{n_0}(R) - \mathbb{G}_{n_1}(R)\right]$$

By assumption 7, we can apply the quantile transformation to yield

$$\mathbb{G}_{n_0}(R) = \alpha_{n_0}(1 - F(R|T = 0))$$

where $\alpha_n(u) = \sqrt{n}(U_n(u) - u)$ and U_n is the empirical distribution of a sample of n independent uniform random variables on $[0, 1]$. Now, $\alpha_{n_0}(1-F(R|T = 0)) = \alpha_{n_0}(1-\hat{F}(R|T = 0) + \hat{F}(R|T = 0) - F(R|T = 0))$, which by definition of the order statistic R is equal to $\alpha_{n_0}(r_n/n_0 + \mathbb{G}_{n_0}(R)/\sqrt{n_0}) = \alpha_{n_0}(r_n/n_0[1 + (\sqrt{n_0}/r_n)\mathbb{G}_{n_0}(R)])$.

By the law of iterated logarithm, $\mathbb{G}_{n_0}(R) = O_{a.s.}(\sqrt{\ln \ln n_0})$, and therefore

$$(\sqrt{n_0}/r_n)\mathbb{G}_{n_0}(R) = O_{\text{a.s.}}(\sqrt{n_0 \ln \ln n_0}/r_n) = o_{\text{a.s.}}(1)$$

under assumption 4. Denoting $u_n := -(\sqrt{n_0}/r_n)\mathbb{G}_{n_0}(R)$, we have $u_n = o_{\text{a.s.}}(1)$ and

$$\sqrt{n_0/r_n}\alpha_{n_0}(r_n/n_0[1 + (\sqrt{n_0}/r_n)\mathbb{G}_{n_0}(R)]) = \sqrt{n_0/r_n}\alpha_{n_0}(r_n/n_0[1 - u_n]).$$

By Mason's central limit theorem for tail empirical processes (see for instance theorem 2.1 page 139 of Einmahl (1992) and del Barrio, Deheuvels, and van de Geer (2007) for a recent account of the theory), there exists a sequence of standard Brownian motions B_n such that

$$\sup_{0<t\leq K<\infty} \left| \sqrt{\frac{n_0}{r_n}}\alpha_{n_0}(\frac{r_n}{n_0}t) - B_n(t) \right| = o_p(1). \tag{5.1}$$

We have therefore $\sqrt{n_0/r_n}\alpha_{n_0}(r_n/n_0[1 - u_n]) = B_n(1 - u_n) + o_p(1)$ and by continuity of Brownian motion sample paths, it follows that $\sqrt{\frac{n_0}{r_n}}\alpha_{n_0}(\frac{r_n}{n_0}(1 - u_n))$ converges weakly to $B(1)$, which is standard normal, and hence so is $\sqrt{\frac{n_0}{r_n}}\mathbb{G}_{n_0}(R)$.

With similarly defined quantities, $\mathbb{G}_{n_1}(R)$ is equal to $\alpha_{n_1}(1 - F(R|T = 1))$. Now:

$$\begin{aligned}
\alpha_{n_1}(1 - F(R|T = 1)) &= \alpha_{n_1}((1 - F(R|T = 0))\xi_R) \\
&= \alpha_{n_1}((1 - \hat{F}(R|T = 0))\xi_R + \xi_R\mathbb{G}_{n_0}(R)/\sqrt{n_0}) \\
&= \alpha_{n_1}\left(\frac{r_n}{n_0}\xi_+ + \frac{r_n}{n_0}(\xi_R - \xi_+)\right. \\
&\quad \left. +\xi_+\mathbb{G}_{n_0}(R)/\sqrt{n_0} + (\xi_R - \xi_+)\mathbb{G}_{n_0}(R)/\sqrt{n_0}\right) \\
&= \alpha_{n_1}\left(\frac{r_n}{n_0}\xi_+\left[1 + \frac{\xi_R - \xi_+}{\xi_+}\right.\right. \\
&\quad \left.\left. +\xi_+\frac{\sqrt{n_0}}{r_n}\mathbb{G}_{n_0}(R) + \frac{\xi_R - \xi_+}{\xi_+}\frac{\sqrt{n_0}}{r_n}\mathbb{G}_{n_0}(R)\right]\right).
\end{aligned}$$

Now

$$\xi_R - \xi_+ = O_p\left(\frac{1 - F(R|T^* = 0)}{1 - F(R|T^* = 1)}\right) = o_p(1).$$

Hence, as previously, $\sqrt{(\rho n_1/(\xi_+ r_n))}\mathbb{G}_{n_1}(R)$ has standard normal limiting distribution.

Finally,

$$\sqrt{r_n}(\hat{\xi}_R - \xi_+) = \xi_+\sqrt{\frac{n_0}{r_n}}\mathbb{G}_{n_0}(R) - \sqrt{\rho\xi_+}\sqrt{\frac{\rho n_1}{\xi_+ r_n}}\mathbb{G}_{n_1}(R) + o_p(1),$$

hence is converges weakly to $\xi_+ Z - \sqrt{\rho\xi_+}Z'$ where Z and Z' are independent standard normal random variables, and, as before, $\rho = \mathrm{P}(T=0)/\mathrm{P}(T=1)$.

To conclude, since $\hat{K}_n = h(\hat{\xi}_R)$, with $h(x) := \frac{1}{1-x}$ a continuous function, the delta method yields

$$\sqrt{r_n}(\hat{K}_n - K) \Rightarrow \frac{1}{(1-\xi_+)^2}(\xi_+ Z - \sqrt{\rho\xi_+}Z').$$

and since $\hat{D} \to_p D$, we have, finally:

$$\sqrt{r_n}(\hat{F}(y|T^*=0) - F(y|T^*=0)) \Rightarrow \frac{D}{(1-\xi_+)^2}(\xi_+ Z - \sqrt{\rho\xi_+}Z').$$

The case of $\hat{F}(y|T^*=1)$ is handled identically, and the asymptotic normality result for the mixture weights estimator follows readily from the delta method. □

It is of practical interest to obtain uniform confidence bands for the proposed nonparametric estimators. To this end, note that the convergence in Theorem 2 is uniform in $y \in \mathbb{R}$, as shown in the proof. Moreover, define

$$\sigma^2(y) := D^2(y)\frac{\xi_+^2 + \rho\xi_+}{(1+\xi_+)^2}, \quad \hat{\sigma}^2(y) := \hat{D}^2(y)\frac{\hat{\xi}_R^2 + \hat{\rho}\hat{\xi}_R}{(1+\hat{\xi}_R)^2},$$

where $D(y) = F(y|T=1) - F(y|T=0)$, $\hat{D}(y) = \hat{F}(y|T=1) - \hat{F}(y|T=0)$ and $\hat{\rho} = n_0/n_1$, then $\sup_{y\in\mathbb{R}}|\hat{\sigma}^2(y) - \sigma^2(y)| = o_p(1)$. From these the next corollary follows immediately.

Corollary 2 *Under assumptions 1-7,*

$$\lim_{n\to\infty} \Pr\left\{\sup_y \left|\frac{\sqrt{r_n}\left(\hat{F}(y|T^*=1) - F(y|T^*=1)\right)}{\hat{\sigma}(y)}\right| > z_{\alpha/2}\right\} = \alpha$$

where z_α denote the $1-\alpha$ quantile of the standard normal distribution.

A similar result holds $F(y|T^*=0)$. One may obtain uniform asymptotic confidence bands for $F(y|T^*=1)$ and $F(y|T^*=0)$ by inverting the inequality in the corollary.

6 Simulations

We now turn to a small Monte Carlo experiment to assess the performance of our proposed estimation procedure. To do this, we set up a mismeasured binary regressor model. The true regressor value $T^* = 0, 1$ is drawn randomly, with $\Pr(T^* = 1) = 0.5$. Following assumption 1, for any given value of T^* the mismeasured regressor T and the outcome Y must be drawn independently of each other. In our experiment, T is drawn randomly conditionally on T^* with transition probabilities

$$p_{11} = \Pr(T = 1|T^* = 1) \text{ and } p_{00} = \Pr(T = 0|T^* = 0).$$

The probabilities p_{00} and p_{11} measure the quality of the measurement of the true regressor; we try values

$$p_{11} = 0.7, 0.8, 0.9 \text{ and } 0.95$$

and set $p_{00} = p_{11}$ in each case. If for instance $p_{11} = p_{00} = 0.8$, then in large samples the regressor will be mismeasured for one observation in five.

We chose to focus on the simplest possible instance of this model: the normal location model, in which outcomes are generated by

$$y_i = (2T_i^* - 1)\beta + u_i$$

where the u_i's are drawn from the standard normal distribution. This specification serves as an unfavorable benchmark for our estimation procedure: Lemma 5 implies that assumption 4 and assumption 6 cannot hold simultaneously in this model. Yet as we will show, the estimator still performs very well in practice.

Given our specification, the unfeasible regression model of y on T^* has a true coefficient equal to 2β, and its R^2 is $\beta^2/(\beta^2 + 1)$; the R^2 of the feasible regression of y on the mismeasured regressor T of course has a lower R^2, especially for smaller values of $p_{00} = p_{11}$. We chose values

$$\beta = 0.25, 0.5, 1 \text{ and } 2,$$

which correspond to "true model" R^2's of $0.06, 0.2, 0.5$ and 0.8.

For each choice of parameters, we simulated $10,000$ samples of size $n = 1,000$ and another 10,000 of size $n = 10,000$. For each such sample, we computed estimators of the cdfs $\hat{F}(y|T^* = 1)$ and $\hat{F}(y|T^* = 0)$ from the formulæ in definition 6, where R is the statistic of order $(r+1)$ of the sample with $T = 0$ and L is the statistic of order $(n-l)$ of the sample of outcomes with $T = 1$. As usual, the asymptotic theory gives little practical guidance

as to optimal values of l and r (and in any case it does not apply to this model.). We experiment with r and l such that $1 - r/n = l/n = 5\%, 10\%, 25\%$ when $n = 1,000$ and $1\%, 5\%, 10\%$ for $n = 10,000$.

Finally, we use the estimated conditional distributions to compute three effects of the true regressor on quantiles: for the median, the upper quartile and the upper decile. We then compare them with the true quantile effects in the model, which are equal to 2β for all quantiles. In the tables, "BIAS" refers to the average estimation error of the quantile effect over the 10,000 replications, and "RMSE" to the root mean squared estimation error. "Decile" refers to the 90%-quantile, "Quartile" to the 75%-quantile, and "Median" of course refers to the 50%-quantile.

Table 1 reports our results when the transition probabilities $p_{00} = p_{11}$ are equal to 0.7 (i.e. 30% of observations are misclassified.) The first thing to notice is that small effects are not properly estimated. This is not surprising: our method is based on tail dominance, and the difference between the tails of two normals with the same variance and means that are so close is very small. Also, one should see these results as a worst-case scenario: other distributions have tails that are better-separated. On the other hand, the results are surprisingly good for $\beta = 1$ and $\beta = 2$. For $\beta = 1$ for instance (a "true model" R^2 of 0.5), and defining tails as 10% of the sample, with as little as 1,000 observations the bias on all three quantile effects is rather small, at about 0.07—recall that the true quantile effects are equal to 2 in this case. The RMSE is about 0.2 for the median, and is somewhat higher for other quantiles as expected.

Next, we investigate the effect of the misclassification probability. Table 2 reports results when transition probabilities are equal to 0.95 (only 5% of observations are misclassified). We find very large overall improvements over the previous case, in which misclassification was much more pronounced. The results for probabilities of 0.8 and 0.9 tell a similar story, and so we do not report them here.

A common concern with estimation methods which depend on a smoothness parameter (here $1 - r/n = l/n$) is the sensitivity of results to this parameter. Our tail estimation procedure is obviously not immune to sensitivity to the choice of order statistic. However, we have explored a very large range (from 1% to 25%); estimation results seem reliable over the whole range, except for the choice 25%, which can be seen to be too extreme for $n = 1,000$.

If the econometrician knew the actual parametric specification of the model, then he could estimate the quantile effect by using maximum likelihood. Table 3 gives the results of such an infeasible benchmark. The weakness of our estimation procedure is apparent

for effects below 0.5; but for larger effects and large samples ($n = 10,000$ rather than $n = 1,000$) it appears to perform quite well, relative to the infeasible alternative.

As the mixture weights $\lambda(T)$ give the probabilities of each regime, their estimation is also of interest, especially in regime switching applications. Here their true values are

$$\lambda(T = 0) = 1 - p_{11} \text{ and } \lambda(T = 1) = p_{11}.$$

We report results for mixture weights estimators in tables 4 and 5. Once again, the RMSEs are large for small values of the regressor β, irrespective of the size of the misclassification error (0.7 or 0.95.) They decrease by about half when the sample size increases from $n = 1000$ to $n = 10,000$. For larger values of the regressor ($\beta = 1$ and $\beta = 2$), the RMSEs tend to be much smaller. Reducing the misclassification error this time has ambiguous effects: it yields better estimates of $\lambda(T = 1)$, but worse estimates of $\lambda(T = 0)$. When using large bandwidths, the RMSEs on $\lambda(T = 1)$ are driven by the bias, whereas the RMSEs on $\lambda(T = 0)$ are driven by the variance. We should note here that getting good estimates of the mixture weights is a hard problem, even with the infeasible maximum likelihood. Table 6 shows the performance of maximum-likelihood estimates of $\lambda(T = 1)$—note that with maximum likelihood we use all parametric assumptions, and so the estimator for $\lambda(T = 0)$ is just a mirror image. Again, when the difference in the location parameters of the components are small ($\beta = 0.25, 0.5$) the maximum likelihood estimates of $\lambda(T = 0)$ and of $\lambda(T = 1)$ have non-negligible biases. Moreover, the estimates often become negative, as indicated by the numbers between brackets: for about 20% of the samples with small β's when the measurement error is large, and even more when it is small since then the true $\lambda(T = 0)$, at 0.05, is closer to zero. We discarded these samples when computing the biases and RMSEs. Surprisingly, our nonparametric estimates are much more robust: none of our samples generated an estimator smaller than zero or larger than one for either mixture weight. We do not have a ready explanation, but we find this to be an appealing property of our method.

Conclusion

We proposed partial identification results under an exclusion restriction that holds for a large variety of mixture models in econometrics, including unobserved heterogeneity models, regime switching models and measurement error models. Partial identification results point naturally to an identification strategy based on tail dominance conditions, which also allow nonparametric estimation of mixtures of two distributions based on intermediate

quantiles. Simulation results show that the nonparametric estimation strategy performs surprisingly well on a Gaussian location model. There are several natural extensions of this work. First, our estimation results should be extended to the case with continuous conditioning information, using conditional quantile methods. Second, partial identification results should be extended to continuous mixtures, to cover a richer class of applications to models with unobserved heterogeneity, in particular nonlinear panel data models and models of games with asymmetric information.

References

Cameron, S., and J. Heckman (1998): "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy*, 106, 262–333.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. Crainiceanu (2006): *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall, CRC Press.

Chamberlain, G. (1986): "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 32, 189–218.

Chen, X., H. Hong, and D. Nekipelov (2009): "Nonlinear models of measurement errors," preprint.

Chen, X., Y. Hu, and A. Lewbel (2008a): "Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information," *Economics Letters*, 100, 381–384.

——— (2008b): "A note on the closed-form identification of regression models with a mismeasured binary regressor," *Statistics and Probability Letters*, 78, 1473–1479.

——— (2009): "Nonparametric identification of regression models containing a misclassified dichotomous regressor without instruments," *Statistica Sinica*, 1, 1–3.

del Barrio, E., P. Deheuvels, and S. van de Geer (2007): *Lectures on Empirical Processes*. European Mathematical Society.

Einmahl, J. (1992): "Limit theorems for tail processes with application to intermediate quantile estimation," *Journal of Statistical Planning and Inference*, 32, 137–145.

Einmahl, U., and D. Mason (1997): "Gaussian approximation of local empirical processes indexed by functions," *Probability Theory and Related Fields*, 107, 283–311.

Farewell, V. (1982): "The use of mixture models for the anaylysis of survival data with long term survivors," *Biometrics*, 38, 1041–1046.

Hall, P., and X.-H. Zhou (2003): "Nonparametric identification of component distributions in a multivariate mixture," *Annals of Statistics*, 31, 201–224.

Hamilton, J. (1989): "A New approach to the analysis of nonstationary times series and the business cycle," *Econometrica*, 57, 357–384.

Heckman, J. (1990): "Variaties of selection bias," *American Economic Review*, 80, 313–318.

Heckman, J., and B. Singer (1984): "Econometric duration analysis," *Journal of Econometrics*, 24, 63–132.

Hendricks, K., J. Pinkse, and R. Porter (2003): "Empirical implications of equilibrium bidding in first-price, symmetric, common value auctions," *Review of Economic Studies*, 70, 115–145.

Horowitz, J., and C. Manski (1995): "Identification and robustness via contaminated and corrupt data," *Econometrica*, 63, 281–302.

Hotz, V., and R. Miller (1993): "Conditional choice probabilities and the estimation of dynamic models," *Review of Economic Studies*, 60, 497–529.

Hu, Y. (2006): "Bounding parameters in a linear regression model with a mismeasured regressor using additional information," *Journal of Econometrics*, 133, 51–70.

Hu, Y. (2008): "Identification and estimation of nonlinear models with misclassification error using instrumental variables," *Journal of Econometrics*, 144, 27–61.

Hu, Y., and M. Shum (2008): "Nonparametric identification of dynamic models with unobserved state variables," unpublished manuscript.

Kasahara, H., and K. Shimotsu (2008): "Nonparametric identification and estimation of multivariate mixtures," Queens Economics Department Working Paper.

Keane, M., and K. Wolpin (1997): "The carreer decisions of young men," *Journal of Political Economy*, 105, 473–522.

Kim, C.-J., and C. Nelson (1998): *State-Space Models with Regime-Switching: Classical and Gibbs-Sampling Approaches with Applications*. MIT Press.

Kitamura, Y. (2003): "Nonprarametric identifiablity of finite mixtures," unpublished manuscript.

Lewbel, A. (2007): "Estimation of average treatment effects with misclassification," *Econometrica*, 75, 537–551.

Lindsay, B., and K. Roeder (1993): "Uniqueness of estimation and identifyability in mixture models," *The Canadian Journal of Statistics*, 21, 139–147.

Mahajan, A. (2006): "Identification and estimation of regression models with misclassification," *Econometrica*, 74, 631–665.

Rootzén, H. (2009): "Weak convergence of the tail empirical process for dependent sequences," *Stochastic Processes and their Applications*, 119, 468–490.

Teicher, H. (1963): "Identifyability of finite mixtures," *Annals of Mathematical Statistics*, 34, 1265–1269.

Table 1: Estimated Quantiles for $p_{11} = p_{00} = 0.7$

β	n	l/n	Decile		Quartile		Median	
			bias	(rmse)	bias	(rmse)	bias	(rmse)
0.25	1000	5%	0.752	(0.961)	0.668	(0.853)	0.635	(0.801)
		10%	0.894	(1.058)	0.783	(0.921)	0.759	(0.876)
		25%	1.221	(1.346)	1.063	(1.137)	0.997	(1.039)
	10,000	1%	0.518	(0.685)	0.459	(0.589)	0.435	(0.545)
		5%	0.590	(0.617)	0.529	(0.554)	0.513	(0.534)
		10%	0.719	(0.733)	0.642	(0.654)	0.621	(0.631)
0.5	1000	5%	0.415	(0.634)	0.350	(0.532)	0.322	(0.478)
		10%	0.457	(0.576)	0.383	(0.472)	0.364	(0.439)
		25%	0.699	(0.748)	0.587	(0.614)	0.569	(0.580)
	10,000	1%	0.207	(0.380)	0.167	(0.294)	0.152	(0.250)
		5%	0.261	(0.289)	0.222	(0.244)	0.216	(0.232)
		10%	0.361	(0.372)	0.309	(0.316)	0.302	(0.308)
1	1000	5%	0.097	(0.589)	0.094	(0.418)	0.090	(0.319)
		10%	0.078	(0.407)	0.067	(0.262)	0.067	(0.200)
		25%	0.161	(0.301)	0.137	(0.197)	0.153	(0.183)
	10,000	1%	0.049	(0.475)	0.0477	(0.303)	0.043	(0.216)
		5%	0.022	(0.197)	0.019	(0.118)	0.020	(0.083)
		10%	0.042	(0.137)	0.034	(0.083)	0.037	(0.063)
2	1000	5%	-0.333	(1.690)	0.026	(0.489)	0.064	(0.317)
		10%	-0.236	(0.893)	0.012	(0.289)	0.029	(0.187)
		25%	-0.082	(0.469)	0.004	(0.158)	0.009	(0.114)
	10,000	1%	-0.252	(0.987)	0.014	(0.342)	0.034	(0.216)
		5%	-0.055	(0.360)	0.002	(0.134)	0.005	(0.063)
		10%	-0.020	(0.204)	0.001	(0.063)	0.002	(0.054)

Table 2: Estimated Quantiles for $p_{11} = p_{00} = 0.95$

β	n	l/n	Decile		Quartile		Median	
			bias	(rmse)	bias	(rmse)	bias	(rmse)
0.25	1000	5%	0.634	(0.699)	0.573	(0.627)	0.558	(0.602)
		10%	0.738	(0.769)	0.667	(0.691)	0.657	(0.677)
		25%	1.058	(1.078)	0.947	(0.957)	0.931	(0.935)
	10,000	1%	0.375	(0.398)	0.347	(0.363)	0.341	(0.353)
		5%	0.544	(0.548)	0.501	(0.504)	0.498	(0.500)
		10%	0.685	(0.687)	0.626	(0.627)	0.621	(0.622)
0.5	1000	5%	0.277	(0.327)	0.250	(0.284)	0.258	(0.281)
		10%	0.373	(0.397)	0.337	(0.354)	0.353	(0.365)
		25%	0.636	(0.650)	0.578	(0.584)	0.606	(0.611)
	10,000	1%	0.137	(0.164)	0.125	(0.141)	0.129	(0.137)
		5%	0.252	(0.256)	0.232	(0.234)	0.246	(0.246)
		10%	0.355	(0.356)	0.326	(0.328)	0.345	(0.346)
1	1000	5%	0.040	(0.197)	0.047	(0.134)	0.066	(0.114)
		10%	0.079	(0.167)	0.091	(0.137)	0.128	(0.154)
		25%	0.214	(0.248)	0.239	(0.245)	0.323	(0.333)
	10,000	1%	0.009	(0.126)	0.010	(0.070)	0.013	(0.044)
		5%	0.038	(0.070)	0.044	(0.054)	0.062	(0.063)
		10%	0.078	(0.089)	0.089	(0.094)	0.125	(0.126)
2	1000	5%	-0.021	(0.248)	0.005	(0.134)	0.009	(0.094)
		10%	0.015	(0.181)	0.041	(0.114)	0.067	(0.109)
		25%	0.128	(0.189)	0.176	(0.200)	0.266	(0.277)
	10,000	1%	-0.014	(0.170)	-0.0004	(0.077)	0.0006	(0.044)
		5%	-0.0005	(0.077)	0.003	(0.031)	0.006	(0.031)
		10%	0.030	(0.063)	0.043	(0.054)	0.066	(0.070)

Table 3: Estimated Quantiles: Infeasible Maximum Likelihood

$p_{11} = p_{00}$	n	β			
		0.25	0.5	1	2
0.7	1000	0.082	-0.018	-0.005	0.0003
		(0.371)	(0.316)	(0.094)	(0.063)
	10,000	0.011	-0.017	-0.0006	-0.00009
		(0.248)	(0.126)	(0.028)	(0.020)
0.95	1000	0.160	-0.010	-0.001	-0.001
		(0.372)	(0.216)	(0.077)	(0.063)
	10,000	0.005	-0.002	-0.0003	-0.00007
		(0.178)	(0.054)	(0.024)	(0.020)

Table 4: Estimated Mixture Weights: $p_{11} = p_{00} = 0.7$

β	n	l/n	$\lambda(T=0)$		$\lambda(T=1)$	
			bias	(rmse)	bias	(rmse)
0.25	1000	5%	0.114	(0.226)	-0.098	(0.231)
		10%	0.132	(0.223)	-0.110	(0.219)
		25%	0.163	(0.215)	-0.119	(0.188)
	10,000	1%	0.090	(0.172)	-0.090	(0.186)
		5%	0.120	(0.144)	-0.092	(0.125)
		10%	0.137	(0.149)	-0.098	(0.116)
0.5	1000	5%	0.055	(0.153)	-0.047	(0.175)
		10%	0.079	(0.134)	-0.042	(0.133)
		25%	0.130	(0.147)	-0.044	(0.090)
	10,000	1%	0.029	(0.097)	-0.026	(0.116)
		5%	0.057	(0.071)	-0.026	(0.057)
		10%	0.079	(0.085)	-0.031	(0.048)
1	1000	5%	0.007	(0.113)	-0.018	(0.147)
		10%	0.014	(0.074)	-0.008	(0.099)
		25%	0.058	(0.068)	0.009	(0.054)
	10,000	1%	0.001	(0.080)	-0.012	(0.105)
		5%	0.005	(0.033)	-0.001	(0.043)
		10%	0.013	(0.025)	0.000	(0.030)
2	1000	5%	-0.000	(0.110)	-0.018	(0.145)
		10%	0.001	(0.071)	-0.008	(0.097)
		25%	0.001	(0.034)	-0.002	(0.053)
	10,000	1%	0.001	(0.079)	-0.009	(0.102)
		5%	-0.000	(0.032)	-0.002	(0.043)
		10%	-0.000	(0.021)	-0.001	(0.029)

Table 5: Estimated Mixture Weights: $p_{11} = p_{00} = 0.95$

β	n	l/n	$\lambda(T=0)$		$\lambda(T=1)$	
			bias	(rmse)	bias	(rmse)
0.25	1000	5%	0.275	(0.293)	-0.217	(0.254)
		10%	0.312	(0.322)	-0.229	(0.049)
		25%	0.374	(0.378)	-0.255	(0.264)
	10,000	1%	0.215	(0.221)	-0.166	(0.182)
		5%	0.276	(0.278)	-0.200	(0.204)
		10%	0.312	(0.313)	-0.220	(0.222)
0.5	1000	5%	0.153	(0.157)	-0.065	(0.096)
		10%	0.203	(0.206)	-0.080	(0.095)
		25%	0.310	(0.311)	-0.107	(0.113)
	10,000	1%	0.085	(0.087)	-0.037	(0.056)
		5%	0.153	(0.154)	-0.059	(0.062)
		10%	0.204	(0.204)	-0.075	(0.077)
1	1000	5%	0.046	(0.048)	-0.002	(0.045)
		10%	0.093	(0.093)	-0.000	(0.030)
		25%	0.228	(0.228)	0.000	(0.018)
	10,000	1%	0.009	(0.012)	-0.001	(0.032)
		5%	0.046	(0.046)	-0.000	(0.014)
		10%	0.093	(0.093)	0.000	(0.010)
2	1000	5%	0.006	(0.009)	-0.001	(0.045)
		10%	0.051	(0.051)	0.002	(0.030)
		25%	0.200	(0.200)	0.010	(0.019)
	10,000	1%	0.000	(0.008)	-0.001	(0.032)
		5%	0.005	(0.005)	0.000	(0.014)
		10%	0.051	(0.051)	0.003	(0.010)

Table 6: Infeasible Maximum Likelihood: Estimating $\lambda(T = 1)$

$p_{11} = p_{00}$	n	β			
		0.25	0.5	1	2
0.7	1000	[22.0%]	[8.1%]	[0.0%]	[0.0%]
		0.048	-0.003	-0.001	-0.000
		(0.118)	(0.075)	(0.030)	(0.021)
	10,000	[15.8%]	[0.3%]	[0.0%]	[0.0%]
		-0.001	-0.007	0.000	-0.000
		(0.096)	(0.036)	(0.009)	(0.007)
0.95	1000	[29.6%]	[28.6%]	[1.2%]	[0.0%]
		0.163	0.037	-0.001	-0.000
		(0.186)	(0.062)	(0.019)	(0.010)
	10,000	[36.0%]	[5.9%]	[0.0%]	[0.0%]
		0.074	0.002	-0.000	0.000
		(0.099)	(0.024)	(0.006)	(0.003)