Identifying Finite Mixtures in Econometric Models

Marc Henry\textsuperscript{1}, Yuichi Kitamura\textsuperscript{2} and Bernard Salanié\textsuperscript{3}


\textsuperscript{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.

\textsuperscript{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.

\textsuperscript{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

Many models used in econometrics can be interpreted as mixtures of distributions. It seems 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, 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 new 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 (e.g. in labor economics, Heckman and Singer (1984) Cameron and Heckman (1998)), with density

$$\prod_{i=1}^{n} \int l(y_i | z_i, u_i; \theta) F(du_i | z_i; \theta)$$

where $u_i$ is the unobserved heterogeneity or “type” for observation $i$. Many dynamic structural models use mixtures to incorporate unobserved heterogeneity. More complex models such of dynamic discrete choice with unobserved states (Rust (1994), Keane and Wolpin (1997)) also belong in this category.

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 is well known, regime switching models also are mixture models. Take the Hamilton (1989) model for the growth of US GDP:

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

where the Gaussian disturbance $\varepsilon_t$ has restricted dynamic dependence and $s_t$ is a hidden two-state Markov chain taking values 0 or 1, so that the economy grows faster on average when the economy is in a boom ($a(1) > a(0)$) then the the density of $y_t$ conditional on its history is

$$\frac{1}{\sigma} \phi \left( \frac{y_t - a(s_t)}{\sigma} \right)$$

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 above conditional density then becomes

$$\sum_{s=0,1} \frac{1}{\sigma} \phi \left( \frac{y_t - a(s_t)}{\sigma} \right) \Pr(s_t = s | y_1, \ldots, y_{t-1})$$

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.

As shown by these examples, a very 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 \( z \).\(^1\) Let \( F \) be the cdf of the data.\(^2\) A conditional mixture model is written as

\[ F(y|z) = \int F(y|z; s) F(ds|z). \]

In this formula, we call \( s \) the mixture index, \( F(ds|z) \) the mixture weights, and each \( F(.|.; s) \) a mixture component.

\( F(y|z) \) is directly identified from the data. 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 \( F(y|z; t) \equiv F(y|z) \) rationalizes the data. The problem remains even if we assume that there are only two mixture components and that the analyst knows that.

The case when the distribution of the mixture index has finite support has generated a very 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 by econometricians 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.


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

\(^1\)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.

\(^2\)The model could be written more generally for ant linear functional \( T_F \) of the cdf \( F \), such as one of its moments, say \( E_F G(Y) \) for some function \( G \).
• 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 (CCP) type estimation procedures Horowitz and Manski (1995), which have gained renewed attention in the recent literature of dynamic structural estimation. In a related application, Hu and Shum (2008) show how to identify a Markov chain with some unobserved components.

Many of the above papers rely on an exclusion restriction, other than a few exceptions, and so do our results. 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 \( F(y|z; s) \) only depend 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 \( z \),

\[
F(.|z; s) \equiv F(.|x; s)
\]

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.

These assumptions have a familiar ring. Take the regression model with misclassified binary regressor in which \( s \) is the actual value of the regressor, \( y \) is an outcome variable, and \( z = (x', w')' \) are explanatory variables. Then our exclusion restriction requires that the distribution of outcomes conditional on regressor value \( s \) only depend on \( x \), while the value of the regressor \( s \) itself also depend on \( w \), i.e. we require that

\[
y \perp \perp w|(x, s).
\]

Compare this with the commonly used conditional independence/ignorability assumption in treatment effects models when selection only acts on observables:

\[
(y(0), y(1)) \perp \perp D|X.
\]

The standard approach in such a model is to estimate the propensity score \( \text{Pr}(D = 1|X) \), and then to estimate the average treatment effect \( E(y(1) - y(0)|X) \) by a variant of the
formula

\[ E(y(1)|p(X), D = 1) - E(y(0)|p(X), D = 0) . \]

(see Hirano, Imbens, and Ridder (2003) for the semiparametrically efficient estimator.)

The analog in our setup would begin by estimating the conditional distribution \( F(dw|x, s) \).

But this is clearly a nonstarter, since \( s \) is unobserved! Essentially, we have to work here in a treatment model in which treatment is not observed: we only observe individuals, their characteristics, and outcomes.

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.) We briefly describe how this can be done in section 3.

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”.

Take the treatment analogy again: if the higher value of the index corresponds to relatively good outcomes, or more productive or compliant individuals, or individuals who are more likely to have been treated, then upon observing a very good outcome the analyst would normally find it quite likely that the individual belongs to the high-index population. 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 an asymptotically normal estimator. The
identification strategy has important similarities with traditional identification at infinity, as proposed by Heckman (1990) and further analyzed among others by Andrews and Schafgans (1998), Lewbel (1998) and Khan and Tamer (2009). 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. Asymptotic behaviour is quite different from that of inverse weighted estimators, as the asymptotic theory relies a strong approximation of the tail empirical process, as for tail index estimation, and the asymptotic distribution is Gaussian.

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:

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

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 \( j = 1 \ldots, J \) 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 last mixture component as reference point and define the new unknown quantities

\[
\Delta(y|x) = (F_1(y|x) - F_J(y|x), \ldots, F_{J-1}(y|x) - F_J(y|x))'
\]

and

\[
\lambda(z) = (\lambda_1(z), \ldots, \lambda_{J-1}(z)).
\]

Then we can write

\[
F(y|z) = F_J(y|x) + \Delta(y|x)\lambda(w);
\]

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

\[
F(y|w, x) - F(y|w', x) = \Delta(y|x)\lambda(w, x) - \lambda(w', x)).
\]

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

\[
(y, w, w') \rightarrow F(y|w, x) - F(y|w', x)
\]

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, to which we will return in section 3.

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 \( x \),

\[
F(y|w', x) - F(y|w, x) = (\lambda(w, x) - \lambda(w', x))(F_1(y|x) - F_2(y|x))
\]

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 “regular” set of values of \( x \):

**Definition 1** Let \( \tilde{X} \) be the set of \( x \) such that for some \((w_1, \ldots, w_J)\) 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 \((F(y_i|w_j, x)\) 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 $w \rightarrow F(y|w, x)$ takes at least two different values.

Definition 1 is just a generalization to $J > 2$. Take $J = 3$ for instance and denote

$$r_{jk;i} = \frac{F(y_i|w_j)}{F(y_i|w_k)} - 1,$$

which is a measure of the relative likelihood of $w_j$ vs $w_k$ when the observed $y$ is smaller than $y_i$. Then it is easy to see that definition 1 requires three things:

1. $r_{13;1} \neq 0$
2. $r_{23;2} \neq 0$
3. $\frac{r_{23;1}}{r_{13;1}} \neq \frac{r_{23;2}}{r_{13;2}}$.

Item 1 (resp. 2) requires that there exist a value $y_1$ where $F(.|w_1)$ (resp. $F(.|w_2)$) and $F(.|w_3)$ differ.

Note that definition 1 involves properties of the mixture. It will prove useful in determining necessary conditions for identification that involve only “observables” (i.e. $F(y|z)$ which is directly identified from the data), as will be shown in theorem 1. We now concentrate on identification conditions on the model structure. As will be apparent in the proof of theorem 1 and 1, $\mathcal{X}$ is a subset of the set $\mathcal{X}$ defined as follows.

Definition 2 Let $\mathcal{X}$ be the set of $x$ such that for some $(w_1, \ldots, w_J)$ 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 $\Delta$ with $(i, j)$-th element $F_j(y_i) - F_j(y_i)$ and the $(J - 1) \times (J - 1)$ matrix $\Lambda$ with $(i, j)$-th element $\lambda_i(w_j) - \lambda_i(w_J)$ are invertible.

Note that the set $\mathcal{X}$ 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 $\mathcal{X}$ 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 $(\lambda, \Delta, F_J)$ and define

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

$$\delta(y|x) = (M(x)^{-1})'\Delta(y|x)$$

$$G_J(y|x) = F_J(y|x) - \Delta(y|x)'v(x).$$

It is easy to check that the new vector of functions $(\mu, \delta, G_J)$ 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.

Again, 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 results shows that admissible $(v, M)$ transforms in fact exactly define partial nonparametric identification:

**Theorem 1** If assumption 1 holds, then then for each $x \in X$, the mixture components $F_1, \ldots, F_J$ and the mixture weights $\lambda$ are nonparametrically identified up to an admissible $(v, M)$ transform.

We now turn to a corollary which provides testable sufficient conditions for the identification condition above.

**Corollary 1** If assumptions 1 holds, then $\tilde{X} \subseteq X$ so that for each $x \in \tilde{X}$ the mixture components $F_1, \ldots, F_J$ and the mixture weights $\lambda$ are nonparametrically identified up to an admissible $(v, M)$ transform.

We prove theorem 1 and corollary 1 jointly.

**Proof** Take an $x \in \tilde{X}$ 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 $X$. From now on, we drop $x$ from the
notation. 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 \(x \in \tilde{X}\).

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

Suppose now \(F\) and \(G\) are observationally equivalent and both satisfy assumptions 1 and 2. Hence we have the following for all \(y\) and \(w\).
\[
F(y|w) = F_J(y) + \Delta(y)^t \lambda(w),
\]
\[
G(y|w) = G_J(y) + \delta(y)^t \mu(w),
\]
from which we deduce
\[
F(y|w) - F(y|w_J) = \Delta(y)^t [\lambda(w) - \lambda(w_J)],
\]
\[
G(y|w) - G(y|w_J) = \delta(y)^t [\mu(w) - \mu(w_J)].
\]
Hence
\[
\Delta(y)^t [\lambda(w) - \lambda(w_J)] = \delta(y)^t [\mu(w) - \mu(w_J)], \tag{1.1}
\]
and
\[
\Delta[\lambda(w) - \lambda(w_J)] = \delta[\mu(w) - \mu(w_J)],
\]
with both \(\Delta\) and \(\delta\) invertible. We therefore have
\[
\mu(w) = M[\lambda(w) + v],
\]
with \(M = \delta^{-1} \Delta\) and \(v = \Delta^{-1} \delta \mu(w_J) - \lambda(w_J)\), so that equation (1.1) becomes
\[
\Delta(y)[\lambda(w) - \lambda(w_J)] = \delta(y)M[\lambda(w) - \lambda(w_J)]
\]
which is true for \( w = w_j \), all \( j = 1, \ldots, J - 1 \). But since the matrix \( \Lambda \) is invertible, we can conclude that \( \delta(y) = M^{-1}\Delta(y) \).

Finally, since \( F \) and \( G \) are observationally equivalent, we have

\[
F_J(y) + \Delta(y)^t\lambda(w) = G_J(y) + \delta(y)^t\mu(w),
\]

hence

\[
G_J(y) = F_J(y) + \Delta(y)^t\lambda(w) - \Delta(y)^tM^{-1}M[\lambda(w) + v]
\]

\[
= F_J(y) - \Delta(y)^tv,
\]

hence the result.

This result 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 \( J \)-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 1 fails.

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

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

\[ F(y|z) = \lambda(z)F_1(y|x) + (1 - \lambda(z))F_2(y|x) \]

\[ = \lambda(z)\Delta(y|x) + F_2(y|x), \]

where \( \Delta(y|x) = F_1(y|x) - F_2(y|x) \).

As explained in section 1, with two components \( \mathcal{X} \) is simply the set of values \( x \) such that the function \( w \rightarrow F(\cdot|w,x) \) takes at least two values (in the space of cdfs.) Take such a value \( x \in \mathcal{X} \); then it follows from lemma 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 full 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 \( T = \emptyset \).

We drop explanatory variables 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 maintained throughout the recent survey of nonlinear and nonclassical measurement error in Chen, Hong, and Nekipelov (2009). It is their assumption 2.1, which is also imposed in recent work on identification of treatment effects under treatment misclassification, including Mahajan (2006), Hu (2006), Lewbel (2007) and Chen, Hu, and Lewbel (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 \( T = 0, 1, \emptyset \). Assumption 2 requires that \( F(.|T) \) 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. 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 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.

### 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

\[
\Pr(S_t = 1|S_{t-1} = 1) = P_{11} \\
\Pr(S_t = 0|S_{t-1} = 0) = P_{00}.
\]

In such a model, Assumption 1 is automatically satisfied with the notation \( y = y_t \) 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 1**

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

\[ \frac{1}{1 - P_{00}} f_1(w) \]

Their condition holds in particular if measurement error is symmetric.
where $f_0$ (resp. $f_1$) is the pdf of $y_t$ conditional on $S_t = 0$ (resp. $S_t = 1$.)

**Proof of Lemma 1:**

$$P(S_t = 1 \text{ and } y_{t-1} \leq w) = P(S_t = 1 \text{ and } y_{t-1} \leq w|S_{t-1} = 1)P(S_{t-1} = 1)$$

$$+ P(S_t = 1 \text{ and } y_{t-1} \leq w|S_{t-1} = 0)P(S_{t-1} = 0)$$

$$= P(S_t = 1|S_{t-1} = 1)P(S_{t-1} = 1)P(y_{t-1} \leq w|S_{t-1} = 1)$$

$$+ P(S_t = 1|S_{t-1} = 0)P(S_{t-1} = 0)P(y_{t-1} \leq w|S_{t-1} = 0)$$

$$= P_{11}P(S_{t-1} = 1)F_1(w) + (1 - P_{00})P(S_{t-1} = 0)F_0(w).$$

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

Moreover,

$$P(y_{t-1} \leq w) = P(y_{t-1} \leq w|S_{t-1} = 1)P(S_{t-1} = 1) + P(y_{t-1} \leq w|S_{t-1} = 0)P(S_{t-1} = 0)$$

$$= P(S_{t-1} = 1)F_1(w) + P(S_{t-1} = 0)F_0(w).$$

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

$$P(S_{t-1} = 1) = \frac{1 - P_{00}}{1 - P_{11} + 1 - P_{00}}$$

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

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 $E(y_t) = \mu_{S_t}$ and $\mu_{S_t} = S_t\mu_1 + (1 - S_t)\mu_2$, and stochastic volatility, with $y_t$ i.i.d. conditionally on $Var(y_t) = \sigma_{S_t}^2$, and $\sigma_{S_t}^2 = S_t\sigma_1^2 + (1 - S_t)\sigma_2^2$.

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

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

13
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 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. Of course, lemma 1 needs to be adapted then.

### 2.3 Unobserved heterogeneity

Consider agents of unobserved type \( \theta = 0 \) or \( \theta = 1 \), and the following model, which is fairly typical of models with unobserved heterogeneity:

\[
y = f(\theta, z, \epsilon).
\]

Then Assumption 1 is implied by the following two restrictions:

1. for all \((\theta, x, \epsilon)\),

\[
f(\theta, z, \epsilon) = f(\theta, x, \epsilon);
\]

or equivalently,

\[
y \perp \perp w | (\theta, x, \epsilon).
\]

2. the distribution of \( \epsilon \) is conditionally independent of \( w \):

\[
\epsilon \perp \perp w | (x, \theta),
\]

Assumption 2 then requires that \( w \) give some information on \( \theta \), conditional on \( x \). Many models used in the literature satisfy these assumptions. A mixed logit model of consumer demand for instance would be

\[
y = \arg \max_{k=1,\ldots,m} \left( u_k(z, \theta) + \epsilon_k \right),
\]

with the \( \epsilon \)'s drawn from iid type-I extreme value distributions conditionally on \( z \) and \( \theta \), so that 2 above trivially holds. Part 1 holds if the mean utilities \( u_k \) do not depend on \( w \), and Assumption 2 requires that the distribution of \( \theta \) 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 \( \theta \), as in an auxiliary statistical model

\[
\theta = 1 \text{ iff } \eta > P(z),
\]

with \( \eta \) 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 can be tested, using the overidentification tests which we already mentioned.

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. Thus our result puts those of of Kasahara and Shimotsu (2008) and Kasahara and Shimotsu (2009) in a wider perspective.

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 $D_i(p_i, p_{-i}, \theta)$, where the demand parameter $\theta$ can take on two values $\bar{\theta} > \theta$.

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 $\theta$ is realized.
- The econometrician later observes costs, prices, market shares, and profits of all firms:

\[
\begin{align*}
\tilde{c}_i &= c_i + \epsilon_{ci} \\
\tilde{p}_i &= p_i + \epsilon_{pi} \\
\tilde{D}_i &= D_i(p_i, p_{-i}, \theta) + \epsilon_{D_i} \\
\tilde{\pi}_i &= (p_i - c_i)D_i(p_i, p_{-i}, \theta) + \epsilon_{\pi_i}.
\end{align*}
\]

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

\[
F(\tilde{D}|\tilde{p}, \tilde{c}) = F(\tilde{D}|\tilde{\pi}, \tilde{p}, \tilde{c}, \theta = \bar{\theta}) \Pr(d = \bar{\theta}|\tilde{\pi}, \tilde{p}, \tilde{c}) + F(\tilde{D}|\tilde{\pi}, \tilde{p}, \tilde{c}, \theta = \theta) \Pr(\theta = \theta|\tilde{\pi}, \tilde{p}, \tilde{c}).
\]
Assumption 1 is satisfied under the following conditions:

1. Prices are observed by the econometrician without measurement error: \( \epsilon_{\pi i} \equiv 0 \).

2. The measurement error on market shares \( \epsilon_{D_i} \) and on costs \( \epsilon_{c_i} \) are independent of the measurement error on profits \( \epsilon_{\pi_i} \), conditional on \( \tilde{c}_i \) and \( \theta \).

When these conditions hold, we have the desired structure:

\[
F(\tilde{D} | \tilde{p}, \tilde{\pi}, \tilde{c}) = F(\tilde{D} | \tilde{p}, \tilde{c}, \theta = \theta) \Pr(\theta = \theta | \tilde{\pi}, \tilde{p}, \tilde{c}) + F(\tilde{D} | \tilde{p}, \tilde{c}, \theta = \theta) \Pr(\theta = \theta | \tilde{\pi}, \tilde{p}, \tilde{c}) 
\]

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 \( \tilde{p_i} \) 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 1. 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 \( z = (w, x) \):

\[
\frac{F(y | z) - F(y | w_0(x), x)}{F(y | w_1(x), x) - F(y | w_0(x), x)} = \frac{\lambda(z) - \lambda(w_0(x), x)}{\lambda(w_1(x), x) - \lambda(w_0(x), x)}.
\]  

(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 \( \lambda \) that rationalizes the data can only differ from \( \Lambda \) by an unknown pair \((\phi(x), \psi(x))\) — what 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 \( \lambda \) in \((w_k(x), x)\) according to the following very simple formulae: it follows from the definition of \( \Lambda \) that
\[
\phi(x) = \lambda(w_0(x), x) \quad \text{and} \quad \psi(x) = \lambda(w_1(x), x) - \lambda(w_0(x), x).
\]
Thus \( \phi(x) \geq 0 \), but \( \psi(x) \) may be negative. For the pair \((\phi, \psi)\), hence the \((v, M)\) transform to be admissible, the corresponding \( \lambda \) needs to be a valid probability. It is easy to obtain identification regions for the functions \( \phi \) and \( \psi \). Denote \( \Lambda(x) = \sup_w \Lambda(w, x) \) and \( \underline{\Lambda}(x) = \inf_w \Lambda(w, x) \). Then the following constraints for \( \phi \) and \( \psi \) result from \( 0 \leq \lambda(z) \leq 1 \):
\[
0 \leq \phi(x) + \psi(x)\Lambda(x) \leq 1
\]
\[
0 \leq \phi(x) + \psi(x)\underline{\Lambda}(x) \leq 1.
\]
Equivalently,
\[
-\psi(x)\Lambda(x) \leq \phi(x) \leq 1 - \psi(x)\Lambda(x)
\]
\[
-\psi(x)\underline{\Lambda}(x) \leq \phi(x) \leq 1 - \psi(x)\underline{\Lambda}(x).
\]
and finally
\[
-\min(\psi(x)\Lambda(x), \psi(x)\underline{\Lambda}(x)) \leq \phi(x) \leq 1 - \max(\psi(x)\Lambda(x), \psi(x)\underline{\Lambda}(x)).
\] (3.2)

The inequalities in (3.2) completely define all admissible \((v, M)\) transforms, and therefore they also define the partially identified regions for \( \lambda \). Note that they immediately imply
\[
\max(\psi(x)\Lambda(x), \psi(x)\underline{\Lambda}(x)) - \min(\psi(x)\Lambda(x), \psi(x)\underline{\Lambda}(x)) \leq 1
\]
and therefore
\[
|\psi(x)| \leq \frac{1}{\Lambda(x) - \underline{\Lambda}(x)}.
\]
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 $\Lambda$ that the bounds $\overline{x}(x)$ and $\underline{x}(x)$ will be further apart; then $\psi(x)$ will be constrained to a smaller interval, so that the variations of $\lambda$ 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 $\overline{x} - \underline{x}$, and hence to a smaller identification region.

![Figure 1: The shaded area is the identified region for the pair $(\psi, \phi)$ in a case where $\overline{\Lambda} < 0$ and $\underline{\Lambda} > 1$. In the binary regressor illustration with missing data, we shall see a case where $\overline{\Lambda} < 0$ and $\underline{\Lambda} = 1$. The only remaining case possible is $\overline{\Lambda} = 0$ and $\underline{\Lambda} = 1$, when the region is uninformative.](image-url)
3.2 Identifying the Components

Once we settle on values for \( \phi \) and \( \psi \) that satisfy (3.2) and thus on a \( \lambda \) that rationalizes the data, the mixture components obtain immediately. For all \( x \in X \),

\[
\Delta(y|x) = F_1(y|x) - F_2(y|x) = \frac{F(y|w_1(x), x) - F(y|w_0(x), x)}{\lambda(w_1(x), x) - \lambda(w_0(x), x)}.
\]

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

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

with the function \( \psi \) defined above.

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

\[
F_2(y|x) = F(y|w_0(x), x) - \lambda(w_0(x), x)\Delta(y|x),
\]

\[
F_1(y|x) = \Delta(y|x) + F_2(y|x)
\]

\[
= F(y|w_0(x), x) + [1 - \lambda(w_0(x), x)]\Delta(y|x).
\]

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

\[
F_2(y|x) = F(y|w_0(x), x) - \frac{\phi(x)}{\psi(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)],
\]

\[
F_1(y|x) = F(y|w_0(x), x) + \frac{1 - \phi(x)}{\psi(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)].
\]

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 \( \Delta \) up to scale is that some quantities of interest are in fact point identified without further assumption. For instance,

**Lemma 2** Take any function \( g \) of the outcome \( y \) such that \( \mathbb{E}_{F_1}g - \mathbb{E}_{F_2}g \neq 0 \). Then for any other function \( f \) of the outcome, the ratio

\[
r(x) = \frac{\mathbb{E}_{F_1}f - \mathbb{E}_{F_2}f}{\mathbb{E}_{F_1}g - \mathbb{E}_{F_2}g}
\]

is point identified.

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.
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 \(F_2\).
3.3 Illustration: Mismeasured Binary Regressor

The construction of $\phi$ and $\psi$ requires that $x \in X$, 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 $\lambda$ in any way: $\Lambda$ 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 $(\psi, \phi)$ space with corners $(-1, 1), (0, 1), \text{ and } (1, 0)$. 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 lemma 2 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

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

(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 $F_2$; we cannot go beyond lemma 2 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 $T = \emptyset$ if the individual is not classified.

Then we can define (for instance)

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

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

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

Now the restrictions on $\phi$ and $\psi$ 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 $(\psi, \phi)$ 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

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

and

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

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

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

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

![Diagram](image)

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

## 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 $F_1(y|x)/F_2(y|x) \to_{y \to -\infty} 0$, $\mathcal{V}_+$ the set of values of $x$ such that $(1 - F_2(y|x))/(1 - F_1(y|x)) \to_{y \to +\infty} 0$.

With this definition, we have

Lemma 3 Under assumption 1, for all $x \in \mathcal{X} \cap \mathcal{V}_- \cap \mathcal{V}_+$, $\lambda$, $F_1$ and $F_2$ 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 $F_2(y|x) \neq 0$ (which will hold for $y$ large enough if $x \in \mathcal{V}_- \cap \mathcal{V}_+$)

$$F(y|z) = F_2(y|x) \left(1 + \lambda(z) \left(\frac{F_1(y|x)}{F_2(y|x)} - 1\right)\right)$$

so that

$$\zeta(y, x) := \frac{F(y|w_1(x), x)}{F(y|w_0(x), x)} = \frac{1 + \lambda(w_1(x), x) \left(\frac{F_1(y|x)}{F_2(y|x)} - 1\right)}{1 + \lambda(w_0(x), x) \left(\frac{F_1(y|x)}{F_2(y|x)} - 1\right)}.$$ 

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

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

Similarly,

$$1 - F(y|z) = (1 - F_2(y|x)) \left(1 - \lambda(z) + \lambda(z) \left(\frac{1 - F_1(y|x)}{1 - F_2(y|x)}\right)\right)$$

so that

$$\xi(y, x) := \frac{1 - F(y|w_1(x), x)}{1 - F(y|w_0(x), x)} = \frac{1 - \lambda(w_1(x), x) + \lambda(w_1(x), x) \left(\frac{1 - F_1(y|x)}{1 - F_2(y|x)}\right)}{1 - \lambda(w_0(x), x) + \lambda(w_0(x), x) \left(\frac{1 - F_1(y|x)}{1 - F_2(y|x)}\right)}.$$ 

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

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

23
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 $x \in X$) gives the result:

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

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:

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

In other words, the distribution of outcomes of the treated dominates the right tail and the distribution of outcomes for the non treated dominates the left tail. Note that nothing is required of the rest of the distribution; in particular, we do not need to impose any stochastic dominance ordering as we did in subsection 3.3. Of course, the roles of right and left tail could be reversed in cases where lower values of the outcome variable are more desirable.

If in particular the mismeasurement is “classical” (independent of $T^*$ and of $y|T^*$), and outcomes are normally distributed conditionally on the regressor, 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 $\Phi$ be the cumulative distribution of a standard normal random variable, and $\mu_1 < \mu_2$ two real numbers. Then

$$
\frac{\Phi\left(\frac{y_1 - \mu_1}{\sigma}\right)}{\Phi\left(\frac{y_2 - \mu_2}{\sigma}\right)} \to 0 \text{ as } y \to +\infty \quad \text{and} \quad \frac{\Phi\left(\frac{y_1 - \mu_2}{\sigma}\right)}{\Phi\left(\frac{y_1 - \mu_1}{\sigma}\right)} \to 0 \text{ as } y \to -\infty.
$$

**Proof** By L'Hôpital’s rule, the result follows if the densities have the required limiting ratios, which is verified as follows. Let $\phi$ be the density of a standard normal random variable. Then, with $K = \exp((\mu_2^2 - \mu_1^2)/(2\sigma^2))$, we have $\phi\left(\frac{y_1 - \mu_1}{\sigma}\right)/\phi\left(\frac{y_1 - \mu_2}{\sigma}\right) = K \exp(2y(\mu_1 - \mu_2)/(2\sigma^2))$, which tends to 0 as $y$ tends to $+\infty$. The other case is treated identically.
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 \(V_- \cap V_+\) corresponds to the set of conditioning points where the regime associated
with \(F_1\) dominates the upper tail, hence is the \textit{expansionary} regime, and the regime associated
with \(F_2\) dominates the lower tail, hence the \textit{contraction} regime. Note that this would be
the case, for instance, with \(F_1\) and \(F_2\) 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 \(\mathcal{X} \cap V_+\) of conditioning points, the following lemma shows
that the dominated regime is fully identified.

**Lemma 5** Under assumption 1, for all \(x \in \mathcal{X} \cap V_+\), \(F_2\) is point identified, whereas \(F_1\) and
\(\lambda\) 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 \(x \in \mathcal{X}\))
\[
\tilde{\kappa}(x) = \frac{\Lambda(w_1(x), x) - \xi_+\Lambda(w_0(x), x)}{\xi_+ - 1}
\]
25
Hence, we have

\[
F(y|z) = \lambda(z)\Delta(y|x) + F_2(y|x)
\]

\[
= (\phi(x) + \psi(x)\Lambda(z))\frac{\Delta(y|x)}{\psi(x)} + F_2(y|x)
\]

\[
= \psi(x)(\kappa(x) + \Lambda(z))\frac{\Delta(y|x)}{\psi(x)} + F_2(y|x)
\]

\[
= (\kappa(x) + \Lambda(z))\Delta(y|x) + F_2(y|x)
\]

and \( F_2 \) is point identified.

The stochastic volatility model illustrates the usefulness of lemma 5. 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 of mixture components

We now propose an estimator for the mixture components inspired from the identification strategy. The identification strategy was based on the identifiable quantities \( \zeta_- := \lim_{y \to -\infty} \frac{F(y|w_1(x), w)}{F(y|w_0(x), x)} \) and \( \zeta_+(x) := \lim_{y \to +\infty} \frac{1 - F(y|w_1(x), x)}{1 - F(y|w_0(x), x)}. \) It was shown that

\[
F_2(y|x) = F(y|w_0(x), x) - \frac{\phi(x)}{\psi(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)],
\]

\[
F_1(y|x) = F(y|w_0(x), x) + \frac{1 - \phi(x)}{\psi(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)],
\]

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

\[
F_2(y|x) = F(y|w_0(x), x) + \frac{1}{1 - \zeta_+(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)],
\]

\[
F_1(y|x) = F(y|w_0(x), x) + \frac{1}{1 - \zeta_-(x)}[F(y|w_1(x), x) - F(y|w_0(x), x)].
\]

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

**Definition 4 (Estimators)** Define \( \hat{\zeta}_-(x) = \hat{F}(L|w_1(x), x)/\hat{F}(L|w_0(x), x) \) and \( \hat{\zeta}_+(x) = [1 - \hat{F}(R|w_1(x), x)]/[1 - \hat{F}(R|w_0(x), x)] \), where \( \hat{F} \) is a non parametric estimator of the conditional cumulative distribution of \( y \) given \( x \) and \( w \).
Example 2.1 continued We now specialize our estimation strategy and give its statistical properties in the case of the mismeasured binary regressor, where the outcome distributions of interest are characterized by $F(y|T^*=1)$ and $F(y|T^*=0)$. An iid sample of individuals is available with their reported regressor values and their outcomes.

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 $\hat{\zeta}_- = \hat{F}(L|T=1)/\hat{F}(L|T=0)$ and $\hat{\xi}_+ = [1 - \hat{F}(R|T=1)]/[1 - \hat{F}(R|T=0)]$. 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 $G_{n_j}(y|T=j) = \sqrt{n_j}(\hat{F}(y|T=j) - F(y|T=j))$, $j = 1, 0$.

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

Definition 6 (Nonparametric estimators for the component distributions:)

$$\hat{F}(y|T^*=0) = \hat{F}(y|T=0) + [1 - \hat{F}(R|T=0)]\frac{\hat{F}(y|T=1) - \hat{F}(y|T=0)}{\hat{F}(R|T=1) - \hat{F}(R|T=0)}$$

$$\hat{F}(y|T^*=1) = \hat{F}(y|T=0) - \hat{F}(L|T=0)\frac{\hat{F}(y|T=1) - \hat{F}(y|T=0)}{\hat{F}(L|T=1) - \hat{F}(L|T=0)}.$$  

Notice that for this estimator to make sense, we need some outcomes of reported non treated 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 $L$:  

27
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 in the untreated category \( (T = 0) \), with \( r_n/n_0 \rightarrow 0 \) and \( r_n/\sqrt{n_0 \ln \ln n_0} \rightarrow \infty \) (and same conditions hold for \( l_n \)).

Remark 2 In practice, one would definitely want to choose \( R \) and \( L \) such that strictly more outcomes of reported treated than outcomes of reported non-treated on the right of \( R \) and strictly more outcomes of reported non-treated than outcomes of reported treated on the left of \( L \). In particular, to avoid the estimator 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 \( 1/2 < \Pr(T^* = 0|T = 0) < 1 \).

To obtain estimators of the distributions of outcomes at a point \( y \), we need the treated and untreated cumulative distributions at that point to be distinct.

Assumption 6 For all \( y \in \mathbb{R} \), \( F(y|T^* = 0) > F(y|T^* = 1) \).

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. Naturally, this applies to the case where positive values of the outcome are associated with desirable outcomes, otherwise, the pattern of dominance is reversed. For asymptotic normality of the proposed estimator, we need to assume this dominance holds at a certain rate.

Assumption 7 \( [1 - F(R|T^* = 0)]/[1 - F(R|T^* = 1)] = o_p(1/\sqrt{n}) \), and \( F(L|T^* = 1)/F(L|T^* = 0) = o_p(1/\sqrt{n}) \).

Remark 3 Assumption 7 is relatively easily verified, and it is generic in that it applies to the case where \( F(y|T^* = 1) \) and \( F(y|T^* = 0) \) are normal with different means and identical variance or when they both have fat tails and satisfy the conditions of the following lemma.

Lemma 6 Suppose \( \alpha_{0R} > \alpha_{1R} > 0 \) and \( \alpha_{0L} > \alpha_{1L} > 0 \) and \( c > 0 \) is a generic positive finite constant (i.e. it may be different in different displays). Suppose the distribution of outcomes for the treated has right (resp. left) tail index \( \alpha_{1R} \) (resp. \( \alpha_{1L} \)), namely \( 1 - F(y|T^* = 1) \) and \( 1 - F(y|T^* = 0) \)
1) \(\sim cy^{-\alpha_1 R}\) when \(y\) tends to \(\infty\) (resp. \(F(y|T^* = 1) \sim c(-y)^{-\alpha_0 L}\) as \(y\) tends to \(-\infty\)). Suppose similarly that the distribution of outcomes for the non treated has right (resp. left) tail index \(\alpha_{0R}\) (resp. \(\alpha_{0L}\)), namely \(1 - F(y|T^* = 0) \sim cy^{-\alpha_{0R}}\) when \(y\) tends to \(\infty\) (resp. \(F(y|T^* = 0) \sim c(-y)^{-\alpha_{0L}}\) as \(y\) tends to \(-\infty\)). Then, under assumptions 4 and 5, assumption 7 holds when

\[
    r_n = o\left(\frac{2(\alpha_{0R} - \alpha_{1R})}{\eta_0}\right) \quad \text{and} \quad l_n = o\left(\frac{2(\alpha_{0L} - \alpha_{1L})}{\eta_0}\right).
\]

In particular, when \(\alpha_{0R} = 2\alpha_{1R}\), assumption 7 will be satisfied if \(r_n = o(n_0^{2/3})\), and when \(\alpha_{0R} = 3\alpha_{1R}\), assumption 7 will be satisfied if \(r_n = o(n_0^{4/5})\), and similarly for \(l_n\).

**Proof of lemma 6** First note that \((1 - F(R|T = 0) = \text{Pr}(T^* = 1|T = 0)[1 - F(R|T^* = 1)] + \text{Pr}(T^* = 0|T = 0)[1 - F(R|T^* = 0)] = (1 - \text{Pr}(T^* = 0|T = 0))[1 - F(R|T = 1)](1 + o_p(1))\) under assumption 5 and \(\alpha_{1R} < \alpha_{0R}\).

Now, by assumption 4, \(r_n/n_0 = 1 - \tilde{F}(R|T = 0) = 1 - F(R|T = 0) + \mathbb{G}_{n_0}(R|T = 0)/\sqrt{n_0} = [1 - F(R|T = 0)] + O_{a.s.}(\ln \ln n_0/\sqrt{n_0})\) 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 7 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 cumulative outcome distribution function of both treated and untreated distributions are invertible.

**Assumption 8** Both \(y \to F(y|T^* = 0)\) and \(y \to F(y|T^* = 1)\) are continuous and strictly increasing.

**Remark 4** Notice that assumption 8 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 1-8, the centered and re-scaled estimator \( \sqrt{n} |\hat{F}(y|T^* = 1) - F(y|T^* = 1)| \) (resp. \( \sqrt{n} |\hat{F}(y|T^* = 0) - F(y|T^* = 0)| \)) is asymptotically normal with mean zero and variance \( V \) (resp. \( \hat{V} \)) with

\[
V = (1 + \rho)[F(y|T = 1) - F(y|T = 0)]^2 \xi_+^2 / (1 - \xi_+)^4 \\
\hat{V} = (1 + \rho)[F(y|T = 1) - F(y|T = 0)]^2 \xi_-^2 / (1 - \xi_-)^4,
\]

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

Proof of Theorem 2 In all that follows, the stochastic dominance relations are uniform with respect to \( y \). Consider first \( \hat{F}(y|T^* = 0) - F(y|T^* = 0) \). It can be decomposed into the sum of six terms, namely \( A_i, i = 0, \ldots, 5 \), with

\[
A_0 = \hat{F}(y|T = 0) - F(y|T = 0) \\
A_1 = \left[ \frac{1 - \hat{F}(R|T = 0)}{F(R|T = 1) - F(R|T = 0)} - \frac{1 - F(R|T = 0)}{F(R|T = 1) - F(R|T = 0)} \right] \\
\times (\hat{F}(y|T = 1) - F(y|T = 0)) \\
A_2 = \left[ \frac{1 - \hat{F}(R|T = 0)}{F(R|T = 1) - F(R|T = 0)} - \lim_{\tau \to \infty} \frac{1 - F(R|T = 0)}{F(R|T = 1) - F(R|T = 0)} \right] \\
\times (\hat{F}(y|T = 1) - F(y|T = 0)) \\
A_3 = \left[ \frac{1 - \hat{F}(R|T = 0)}{F(R|T = 1) - F(R|T = 0)} - \frac{1 - F(R|T = 0)}{F(R|T = 1) - F(R|T = 0)} \right] \\
\times (\hat{F}(y|T = 1) - F(y|T = 0) - F(y|T = 1) + F(y|T = 0)) \\
A_4 = \left[ \frac{1 - F(R|T = 0)}{F(R|T = 1) - F(R|T = 0)} - \lim_{\tau \to \infty} \frac{1 - F(R|T = 0)}{F(R|T = 1) - F(R|T = 0)} \right] \\
\times (\hat{F}(y|T = 1) - F(y|T = 0) - F(y|T = 1) + F(y|T = 0)) \\
A_5 = \left[ \lim_{\tau \to \infty} \frac{1 - F(R|T = 0)}{F(R|T = 1) - F(R|T = 0)} \right] \\
\times (\hat{F}(y|T = 1) - F(y|T = 0) - F(y|T = 1) + F(y|T = 0)).
\]

By definition \( A_0 \) is equal to \( G_n(y|T = 0)/\sqrt{n_0} \), where \( G_n(.|T = 0) \) denotes the empirical process relative to the sample of classified untreated individuals.

\( A_5 \) is equal to \( \frac{1}{\xi_+} \times [n_1^{-1/2} G_{n_1}(y|T = 1) - n_0^{-1/2} G_{n_0}(y|T = 0)] \).

\( A_4 = \rho_1(1) \times A_5 \).
\( A_3 \) is
\[
A_3 = A_1 \times \frac{\hat{F}(y|T = 1) - F(y|T = 1) - [\hat{F}(y|T = 0) - F(y|T = 0)]}{F(y|T = 1) - F(y|T = 0)}
\]
\[
= A_1 \times \frac{n_1^{-1/2}G_{n_1}(y|T = 1) - n_0^{-1/2}G_{n_0}(y|T = 0)}{F(y|T = 1) - F(y|T = 0)},
\]
which, by the law of iterated logarithm, is almost surely dominated by
\[
A_1 \times \max \left( \frac{\ln \ln n_1}{\sqrt{n_1}}, \frac{\ln \ln n_0}{\sqrt{n_0}} \right),
\]
hence by \( A_1 \) under assumption 6 and assumption 5. Indeed, \( F(y|T = 0) - F(y|T = 1) = F(y|T^* = 1)[1 - \Pr(T^* = 0|T = 0) - \Pr(T^* = 1|T = 1)] + F(y|T^* = 0)[\Pr(T^* = 0|T = 0) + \Pr(T^* = 1|T = 1) - 1]\) \( \times [F(y|T^* = 0) - F(y|T^* = 1)]. \) Both terms in brackets are non zero by assumption 5 and assumption 6 respectively. We have shown, therefore, that \( A_3 \) is dominated by \( A_1 \) and that
\[
\sqrt{n_0}(A_0 + A_4 + A_5) = \frac{1}{1 - \xi_+} \left( \sqrt{\frac{n_0}{n_1}}G_{n_1}(y|T = 1) - \xi_+ G_{n_0}(y|T = 0) \right) (1 + o_p(1)) \quad (5.1)
\]
so that it converges weakly to a centered normal distribution with variance \( V_0 \) as follows, with \( \rho = \Pr(T = 0)/\Pr(T = 1). \)
\[
V_0 = \frac{1}{(1 - \xi_+)^2} \left\{ \rho F(y|T = 1)[1 - F(y|T = 1)] + \xi_+^2 F(y|T = 1)[1 - F(y|T = 1)] \right\}. \quad (5.2)
\]
Consider now \( A_2. \) Define \( \xi_R = [1 - F(R|T = 1)]/[1 - F(R|T = 0)]. \) We have
\[
A_2 = \frac{1 - F(R|T = 0)}{F(R|T = 1) - F(R|T = 0)} - \frac{1}{1 - \xi_+}
\]
\[
= \frac{1}{\frac{1 - F(R|T = 1)}{1 - F(R|T = 0)} - \frac{1}{1 - \xi_+} - \frac{1}{1 - \xi_R}} = \frac{\xi_R - \xi_+}{(1 - \xi_R)(1 - \xi_+)},
\]
which is \((1 - \xi_+)^{-2}(\xi_R - \xi_+)(1 + o_p(1)). \) So we have
\[
\xi_R = \frac{\Pr(T^* = 1|T = 1)[1 - F(R|T^* = 1)] + (1 - \Pr(T^* = 1|T = 1))[1 - F(R|T^* = 0)]}{(1 - \Pr(T^* = 0|T = 0))[1 - F(R|T^* = 1)] + \Pr(T^* = 0|T = 0)[1 - F(R|T^* = 0)]},
\]
\[
= \frac{\Pr(T^* = 1|T = 1) + [1 - \Pr(T^* = 1|T = 1)] \frac{1 - F(R|T^* = 0)}{1 - F(R|T^* = 1)}}{[1 - \Pr(T^* = 0|T = 0)] + \Pr(T^* = 0|T = 0) \frac{1 - F(R|T^* = 0)}{1 - F(R|T^* = 1)}}.
\]
Hence $\xi_R - \xi_+$ is equal to

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

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 $A_2$ is $O_p([1 - F(R|T^* = 0)]/[1 - F(R|T^* = 1)])$, which is $o_p(1/\sqrt{n})$ by assumption 7.

Now consider $A_1$. $A_1$ is equal to $[F(y|T = 1) - F(y|T = 0)] \times A'_1$ with $A'_1 = (1 - \hat{\xi}_R)^{-1} - (1 - \xi_R)^{-1}$, and $\hat{\xi}_R = [1 - \hat{F}(R|T = 1)]/[1 - \hat{F}(R|T = 0)]$. We have

$$
A'_1 = \frac{\hat{\xi}_R - \xi_R}{(1 - \hat{\xi}_R)(1 - \xi_R)}
= \frac{\hat{\xi}_R - \xi_R}{(1 - \hat{\xi}_R)^2} + \hat{\xi}_R - \xi_R \times \left[ \frac{1}{(1 - \xi_R)(1 - \xi_R)} - \frac{1}{(1 - \hat{\xi}_R)^2} \right]
= \frac{\hat{\xi}_R - \xi_R}{(1 - \hat{\xi}_R)^2} + (\hat{\xi}_R - \xi_R) \times o_p(1).
$$

By definition of $R$, since $\hat{F}$ is the empirical distribution relative to the sample of indi-
individually classified as non treated, we have $1 - \hat{F}(R|T = 0) = r_n/n_0$. Hence,

$$
\hat{\xi}_R - \xi_R
= \frac{\hat{\xi}_R - 1 - F(R|T = 1)}{1 - F(R|T = 0)} + [1 - F(R|T = 1)] \left( \frac{1}{1 - \hat{F}(R|T = 0)} - 1 \right)
= \frac{\mathbb{G}_{n_0}(R|T = 0)}{\sqrt{n_0}} + [1 - F(R|T = 1)] \frac{\mathbb{G}_{n_0}(R|T = 0)}{\sqrt{n_0}} (1 - F(R|T = 0))
= \frac{\sqrt{n_0}}{r_n} \left( \xi_R \mathbb{G}_{n_0}(R|T = 0) - \sqrt{\frac{n_0}{n_1}} \mathbb{G}_{n_1}(R|T = 1) \right)
= \frac{\sqrt{n_0}}{r_n} \left( \xi_{+} \mathbb{G}_{n_0}(R|T = 0) - \sqrt{\frac{n_0}{n_1}} \mathbb{G}_{n_1}(R|T = 1) \right) + \frac{\sqrt{n_0}}{r_n} (\xi_R - \xi_{+}) \mathbb{G}_{n_0}(R|T = 0).
$$

Since $(\xi_R - \xi_{+}) = O_p([1 - F(R|T^* = 0)]/[1 - F(R|T^* = 1)])$, it suffices to consider the first term. Define

$$
A'_2 = \frac{\sqrt{n_0}}{r_n} \left( \xi_{+} \mathbb{G}_{n_0}(R|T = 0) - \sqrt{\frac{n_0}{n_1}} \mathbb{G}_{n_1}(R|T = 1) \right).
$$

By assumption 8, we can apply the quantile transformation to yield

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

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 $\alpha_{n_0}(r_n/n_0 + \mathbb{G}_{n_0}(R|T = 0)/\sqrt{n_0}) = \alpha_{n_0}(r_n/n_0[1 + (\sqrt{n_0}/r_n)] \mathbb{G}_{n_0}(R|T = 0)))$.

By the law of iterated logarithm, $\mathbb{G}_{n_0}(R|T = 0) = O_{a.s.}(\ln \ln n_0)$, and therefore $(\sqrt{n_0}/r_n) \mathbb{G}_{n_0}(R|T = 0) = O_{a.s.}(\sqrt{n_0}/r_n) = o_{a.s.}(1)$ under assumption 4.

By Mason’s strong approximation theorem for tail empirical processes (see for instance proposition 3.10 page 156 of del Barrio, Deheuvels, and van de Geer (2007)), there exists a sequence of standard Brownian motions $B_n$ such that

$$
\sup_{0 \leq s \leq K < \infty} \left| \sqrt{\frac{n_0}{r_n}} \left( \alpha_{n_0}(\frac{r_n}{n_0} s) - B_n(\frac{r_n}{n_0} s) \right) \right| = o_{a.s.}(1).
$$

(5.3)

Since $\sqrt{\frac{n_0}{r_n}} B_n(\frac{r_n}{n_0} s)$ is a standard Brownian motion, it follows that $\sqrt{\frac{n_0}{r_n}} \alpha_{n_0}(\frac{r_n}{n_0} (1 + o_{a.s.}(1)))$ converges weakly to $B(1)$, which is standard normal.
With similarly defined quantities, \( \mathbb{G}_{n1}(R|T = 1) \) is equal to \( \alpha_{n1}(1 - F(R|T = 1)) \). Now:

\[
\alpha_{n1}(1 - F(R|T = 1)) = \alpha_{n1}((1 - F(R|T = 0))\xi_R) = \alpha_{n1}((1 - \hat{F}(R|T = 0))\xi_R + \xi_R\mathbb{G}_{n0}(R|T = 0)/\sqrt{n_0}) = \alpha_{n1}\left(\frac{r_n}{n_0}\xi_R + \frac{r_n}{n_0}(\xi_R - \xi_+) + \xi_+\mathbb{G}_{n0}(R|T = 0)/\sqrt{n_0} + (\xi_R - \xi_+)\mathbb{G}_{n0}(R|T = 0)/\sqrt{n_0}\right).
\]

Since \( \xi_R - \xi_+ = o_p(1) \) and \( (\sqrt{n_0}/r_n)\mathbb{G}_{n0}(R|T = 0) = o_{a.s.}(1) \), as explained previously, \( \sqrt{(n_0/(\xi_++r_n))}\mathbb{G}_{n1}(R|T = 1) \) has standard normal limiting distribution.

Finally, the distribution of \( \sqrt{n}A_1 \) converges to that of

\[
[F(y|T = 1) - F(y|T = 0)] \frac{\xi_+}{(1 - \xi_+)^2} (Z - \sqrt{\rho}Z'),
\]

where \( Z \) and \( Z' \) are independent standard normal random variables, i.e. \( \sqrt{n}A_1 \Rightarrow N(0, V_1) \), with

\[
V_1 = (1 + \rho)[F(y|T = 1) - F(y|T = 0)]^2\xi_+^2/(1 - \xi_+)^4,
\]

where, as before, \( \rho = P(T = 0)/P(T = 1) \).

The asymptotic analysis of the untreated distribution estimator is almost identical. It is developed below for completeness. \([\hat{F}(y|T^* = 1) - F(y|T^* = 1)]\) can be decomposed in the same way into \( A_0 \) and five additional terms \( \tilde{A}_1 \) to \( \tilde{A}_5 \).

\[
\tilde{A}_1 = \left[\frac{-\hat{F}(L|T = 0)}{\hat{F}(L|T = 1) - \hat{F}(L|T = 0)} - \frac{-F(L|T = 0)}{F(L|T = 1) - F(L|T = 0)}\right] \times (F(y|T = 1) - F(y|T = 0));
\]

\[
\tilde{A}_2 = \left[\frac{-F(L|T = 0)}{F(L|T = 1) - F(L|T = 0)} - \lim_{R \to -\infty} \frac{F(L|T = 0)}{F(L|T = 1) - F(L|T = 0)}\right] \times (F(y|T = 1) - F(y|T = 0));
\]

\[
\tilde{A}_3 = \left[\frac{-\hat{F}(L|T = 0)}{\hat{F}(L|T = 1) - \hat{F}(L|T = 0)} - \frac{-F(L|T = 0)}{F(L|T = 1) - F(L|T = 0)}\right] \times (\hat{F}(y|T = 1) - \hat{F}(y|T = 0) - F(y|T = 1) + F(y|T = 0));
\]

\[
\tilde{A}_4 = \left[\frac{-F(L|T = 0)}{F(L|T = 1) - F(L|T = 0)} - \lim_{R \to -\infty} \frac{F(L|T = 0)}{F(L|T = 1) - F(L|T = 0)}\right] \times (\hat{F}(y|T = 1) - \hat{F}(y|T = 0) - F(y|T = 1) + F(y|T = 0));
\]

\[
\tilde{A}_5 = \left[\lim_{R \to -\infty} \frac{F(L|T = 1) - F(L|T = 0)}{F(L|T = 1) - F(L|T = 0)}\right] \times (\hat{F}(y|T = 1) - \hat{F}(y|T = 0) - F(y|T = 1) + F(y|T = 0)).
\]
As before, we have \( \tilde{A}_3 = o_p(\tilde{A}_1) \) and
\[
\sqrt{n_0}(A_0 + \tilde{A}_4 + \tilde{A}_5) = \frac{1}{1 - \zeta} \left( \sqrt{\frac{n_0}{n_1}} g_{n_1}(y|T = 1) - \zeta G_{n_0}(y|T = 0) \right) + o_p(1) \tag{5.5}
\]
so that it converges weakly to a centered normal distribution with variance \( \tilde{V}_0 \) as follows, with \( \rho = \lim n_0/n_1 \) as \( n \to \infty \).

\[
\tilde{V}_0 = \frac{1}{(1 - \zeta)^2} \left\{ \rho F(y|T = 1)\{1 - F(y|T = 1)\} + \zeta^2 F(y|T = 1)\{1 - F(y|T = 1)\} \right\}. \tag{5.6}
\]

Consider now \( \tilde{A}_2 \). We have
\[
\tilde{A}_2 = \frac{-F(L|T = 0)}{F(L|T = 1) - F(L|T = 0)} - \frac{1}{1 - \zeta} = \frac{1}{1 - \frac{F(L|T = 1)}{F(L|T = 0)}} - \frac{1}{1 - \zeta} = \frac{\zeta_L - \zeta}{(1 - \zeta_L)(1 - \zeta)},
\]
which is \((\zeta_L - \zeta)(1 + o(1))\), with \( \zeta_L = F(L|T = 1)/F(R|T = 0). \) So we have
\[
\zeta_L = \frac{\Pr(T^* = 1|T = 1)F(L|T^* = 1) + (1 - \Pr(T^* = 1|T = 1))F(L|T^* = 0)}{(1 - \Pr(T^* = 0|T = 0))F(L|T^* = 1) + \Pr(T^* = 0|T = 0)F(L|T^* = 0)},
\]
\[
= \frac{\Pr(T^* = 1|T = 1)F(L|T^* = 1)F(L|T^* = 1) + [1 - \Pr(T^* = 1|T = 1)]}{[1 - \Pr(T^* = 0|T = 0)]F(L|T^* = 1) + \Pr(T^* = 0|T = 0)}.
\]

Hence \( \zeta_L - \zeta \) is equal to
\[
\zeta_L - \frac{1 - \Pr(T^* = 1|T = 1)}{\Pr(T^* = 0|T = 0)} = \frac{\Pr(T^* = 0|T = 0)\left( \Pr(T^* = 1|T = 1)F(L|T^* = 1) + [1 - \Pr(T^* = 1|T = 1)] \right)}{\Pr(T^* = 0|T = 0)\left( [1 - \Pr(T^* = 0|T = 0)] F(L|T^* = 1) + \Pr(T^* = 0|T = 0) \right)}
\]
\[
= \frac{[\Pr(T^* = 0|T = 0) + \Pr(T^* = 1|T = 1) - 1] F(L|T^* = 1)}{[\Pr(T^* = 0|T = 0)] F(L|T^* = 1) + \Pr(T^* = 0|T = 0)},
\]
which is \( O(F(L|T^* = 1)/F(L|T^* = 0)) \) because \( 0 < \Pr(T^* = 0|T = 0) < 1 \) under assumption 5. Hence we have shown that \( \tilde{A}_2 \) is \( o_p(F(L|T^* = 1)/F(L|T^* = 0)) \), which is \( o_p(1/\sqrt{n_0}) \) by assumption 7.
Finally, consider $\tilde{A}_1$. $\tilde{A}_1$ is equal to $[F(y|T = 1) - F(y|T = 0)] \times \tilde{A}_1'$ with $\tilde{A}_1' = (1 - \tilde{\zeta})^{-1} - (1 - \zeta_L)^{-1}$, and $\tilde{\zeta} = \hat{F}(L|T = 1)/\hat{F}(L|T = 0)$. We have

$$\tilde{A}_1' = \frac{\hat{\zeta}_L - \zeta_L}{(1 - \tilde{\zeta}_L)(1 - \zeta_L)}$$

$$= \frac{\hat{\zeta}_L - \zeta_L}{(1 - \tilde{\zeta}_L)^2} + (\tilde{\zeta}_L - \zeta_L) \times \left[ \frac{1}{(1 - \tilde{\zeta}_L)(1 - \zeta_L)} - \frac{1}{(1 - \zeta_L)^2} \right]$$

$$= \frac{\hat{\zeta}_L - \zeta_L}{(1 - \tilde{\zeta}_L)^2} + (\tilde{\zeta}_L - \zeta_L) \times \left[ \frac{(\hat{\zeta}_L - \zeta_L) \times o(1) + o(1)}{(1 - \tilde{\zeta}_L)^4 + (\hat{\zeta}_L - \zeta_L) \times o(1) + o(1)} \right]$$

$$= \frac{\hat{\zeta}_L - \zeta_L}{(1 - \tilde{\zeta}_L)^2} + (\tilde{\zeta}_L - \zeta_L) \times o(1).$$

By definition of $L$, since $\hat{F}$ is the empirical distribution relative to the sample of individuals classified as non treated, we have $\hat{F}(L|T = 0) = l_n/n_0$. Hence,

$$\hat{\zeta}_L - \zeta_L$$

$$= \hat{\zeta}_L - \frac{F(L|T = 1)}{F(L|T = 0)} + [F(L|T = 1)] \left( \frac{1}{\hat{F}(L|T = 0)} - \frac{1}{F(L|T = 0)} \right)$$

$$= \frac{G_{n_1}(L|T = 1)}{\sqrt{n_1}} \frac{l_n}{n_0} + [F(L|T = 1)] \frac{-G_{n_0}(L|T = 0)}{\sqrt{n_0}} \frac{l_n}{n_0} \frac{F(L|T = 0)}{n_0}$$

$$= \frac{G_{n_1}(L|T = 1)}{\sqrt{n_1}} - \frac{F(L|T = 1)}{\hat{F}(L|T = 0)} \frac{G_{n_0}(L|T = 0)}{\sqrt{n_0}} \frac{l_n}{n_0}$$

$$= \frac{\sqrt{n_0}}{l_n} \left( \sqrt{\frac{n_0}{n_1}} G_{n_1}(L|T = 1) - \zeta_L G_{n_0}(L|T = 0) \right)$$

$$= \frac{\sqrt{n_0}}{l_n} \left( \sqrt{\frac{n_0}{n_1}} G_{n_1}(L|T = 1) - \zeta L G_{n_0}(L|T = 0) \right) + \frac{\sqrt{n_0}}{l_n} ((\zeta_L - \zeta) G_{n_0}(L|T = 0)).$$

Since $(\zeta_L - \zeta) = O_p(F(L|T^* = 1)/F(L|T^* = 0))$, it suffices to consider the first term. Define

$$\tilde{A}_1'' = \frac{\sqrt{n_0}}{l_n} \left( \sqrt{\frac{n_0}{n_1}} G_{n_1}(L|T = 1) - \zeta - G_{n_0}(L|T = 0) \right).$$

As before, we have $G_{n_0}(L|T = 0) = \alpha_{n_0}(F(L|T = 0))$. By definition of $L$, the latter is equal to $\alpha_{n_0}(\hat{F}(L|T = 0) + F(L|T = 0) - \hat{F}(L|T = 0)) = \alpha_{n_0}(l_n/n_0 - G_{n_0}(L|T = 0)/\sqrt{n_0})$. By the law of iterated logarithm, $G_{n_0}(L|T = 0) = O_{a.s.}( \ln \ln n_0 )$, and therefore $\sqrt{(n_0/l_n)} \alpha_{n_0}(l_n/n_0 + G_{n_0}(L|T = 0)/\sqrt{n_0})$ has the same limiting distribution as $\sqrt{(n_0/l_n)} \alpha_{n_0}(l_n/n_0)$ which is standard normal by Mason’s strong approximation theorem for tail empirical processes.
Similarly, $G_{n_1}(L|T = 1)$ has the same distribution as $\alpha_{n_1}(F(L|T = 1))$. Now:

$$\begin{align*}
\alpha_{n_1}(F(L|T = 1)) &= \alpha_{n_1}((F(L|T = 0))\zeta_L) \\
&= \alpha_{n_1}((\hat{F}(L|T = 0))\zeta_L - \zeta_L G_{n_0}(L|T = 0)/\sqrt{n_0}) \\
&= \alpha_{n_1}\left(\frac{l_0}{n_0}\zeta_- + \frac{l_n}{n_0}(\zeta_L - \zeta_-) - \zeta_- G_{n_0}(L|T = 0)/\sqrt{n_0} + (\zeta_- - \zeta_L) G_{n_0}(L|T = 0)/\sqrt{n_0}\right).
\end{align*}$$

Hence, as argued before, $\sqrt{(n_0/\zeta_- x l_n)}G_{n_1}(L|T = 1)$ has the same limiting distribution as $\sqrt{(n_0/\zeta_- x l_n)}\alpha_{n_1}(\zeta_- x l_n/n_0)$, which is also standard normal.

Finally, the distribution of $\sqrt{l_n}\tilde{A}_1$ converges to that of

$$[F(y|T = 1) - F(y|T = 0)]\frac{\zeta_-}{(1 - \zeta_-)^2} (\sqrt{p}Z - Z'),$$

where $Z$ and $Z'$ are independent standard normal random variables, i.e. $\sqrt{l_n}\tilde{A}_1 \Rightarrow N\left(0, V_1\right)$, with

$$V_1 = (1 + \rho)^2|F(y|T = 1) - F(y|T = 0)|^2\zeta_-^2/(1 - \zeta_-)^4. \tag{5.7}$$

6 Simulations

We now turn to a small Monte Carlo experiment to assess the performance of our proposed estimation procedure in the point identified case. To do this, we set up a mismeasured binary regressor model. The true regressor value $T^* = 0, 1$ is drawn randomly with probability 0.5 of $T^* = 1$. Following assumption 1, for any given value of $T^*$ the mismeasured regressor $\tilde{T}$ and the outcome $Y$ must be drawn independently of each other. In our experiment, $T$ is drawn randomly conditionally on $T^*$ with transition probabilities $p_{11} = \Pr(T = 1|T^* = 1)$ and $p_{00} = \Pr(T = 0|T^* = 0)$. The probabilities $p_{00}$ and $p_{11}$ measure the quality of the measurement of 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.

To satisfy assumption 7, outcomes $Y$ are drawn independently, conditionally on $T^*$, from a normal distribution with mean $\beta(2T^* - 1)$ and unit variance. Thus the true regression coefficient is $2\beta$, and the $R^2$ of the true, unfeasible regression model is $\beta^2/(\beta^2 + 1)$; the $R^2$
of the feasible model, estimated with the mismeasured regressor, 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 formulae 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 theory gives little practical guidance as to optimal values of $l$ and $r$. 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 $2\beta$ 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 unexpected: 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, of 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 of course somewhat higher for more extreme quantiles.

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 be a reasonable robust alternative.
Table 1: $p_{11} = p_{00} = 0.7$

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$n$</th>
<th>$l/n$</th>
<th>Decile</th>
<th>Quartile</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>bias (rmse)</td>
<td>bias (rmse)</td>
<td>bias (rmse)</td>
</tr>
<tr>
<td>0.25</td>
<td>5%</td>
<td>1000</td>
<td>0.752 (0.961)</td>
<td>0.668 (0.853)</td>
<td>0.635 (0.801)</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>1,000</td>
<td>0.894 (1.058)</td>
<td>0.783 (0.921)</td>
<td>0.759 (0.876)</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td></td>
<td>1.221 (1.346)</td>
<td>1.063 (1.137)</td>
<td>0.997 (1.039)</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>10,000</td>
<td>0.518 (0.685)</td>
<td>0.459 (0.589)</td>
<td>0.435 (0.545)</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td></td>
<td>0.590 (0.617)</td>
<td>0.529 (0.554)</td>
<td>0.513 (0.534)</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td></td>
<td>0.719 (0.733)</td>
<td>0.642 (0.654)</td>
<td>0.621 (0.631)</td>
</tr>
<tr>
<td>0.5</td>
<td>5%</td>
<td>1000</td>
<td>0.415 (0.634)</td>
<td>0.350 (0.532)</td>
<td>0.322 (0.478)</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td></td>
<td>0.457 (0.576)</td>
<td>0.383 (0.472)</td>
<td>0.364 (0.439)</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td></td>
<td>0.699 (0.748)</td>
<td>0.587 (0.614)</td>
<td>0.569 (0.580)</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>10,000</td>
<td>0.207 (0.380)</td>
<td>0.167 (0.294)</td>
<td>0.152 (0.250)</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td></td>
<td>0.261 (0.289)</td>
<td>0.222 (0.244)</td>
<td>0.216 (0.232)</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td></td>
<td>0.361 (0.372)</td>
<td>0.309 (0.316)</td>
<td>0.302 (0.308)</td>
</tr>
<tr>
<td>1</td>
<td>5%</td>
<td>1000</td>
<td>0.097 (0.589)</td>
<td>0.094 (0.418)</td>
<td>0.090 (0.319)</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td></td>
<td>0.078 (0.407)</td>
<td>0.067 (0.262)</td>
<td>0.067 (0.200)</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td></td>
<td>0.161 (0.301)</td>
<td>0.137 (0.197)</td>
<td>0.153 (0.183)</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>10,000</td>
<td>0.049 (0.475)</td>
<td>0.0477 (0.303)</td>
<td>0.043 (0.216)</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td></td>
<td>0.022 (0.197)</td>
<td>0.019 (0.118)</td>
<td>0.020 (0.083)</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td></td>
<td>0.042 (0.137)</td>
<td>0.034 (0.083)</td>
<td>0.037 (0.063)</td>
</tr>
<tr>
<td>2</td>
<td>5%</td>
<td>1000</td>
<td>-0.333 (1.690)</td>
<td>0.026 (0.489)</td>
<td>0.064 (0.317)</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td></td>
<td>-0.236 (0.893)</td>
<td>0.012 (0.289)</td>
<td>0.029 (0.187)</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td></td>
<td>-0.082 (0.469)</td>
<td>0.004 (0.158)</td>
<td>0.009 (0.114)</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>10,000</td>
<td>-0.252 (0.987)</td>
<td>0.014 (0.342)</td>
<td>0.034 (0.216)</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td></td>
<td>-0.055 (0.360)</td>
<td>0.002 (0.134)</td>
<td>0.005 (0.063)</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td></td>
<td>-0.020 (0.204)</td>
<td>0.001 (0.063)</td>
<td>0.002 (0.054)</td>
</tr>
</tbody>
</table>
Table 2: $p_{11} = p_{00} = 0.95$

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$n$</th>
<th>$l/n$</th>
<th>Decile</th>
<th>Quartile</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>bias (rmse)</td>
<td>bias (rmse)</td>
<td>bias (rmse)</td>
</tr>
<tr>
<td>0.25</td>
<td>5%</td>
<td>0.634 (0.699)</td>
<td>0.573 (0.627)</td>
<td>0.558 (0.602)</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.738 (0.769)</td>
<td>0.667 (0.691)</td>
<td>0.657 (0.677)</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>1.058 (1.078)</td>
<td>0.947 (0.957)</td>
<td>0.931 (0.935)</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>5%</td>
<td>0.277 (0.327)</td>
<td>0.250 (0.284)</td>
<td>0.258 (0.281)</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.373 (0.397)</td>
<td>0.337 (0.354)</td>
<td>0.353 (0.365)</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>0.636 (0.650)</td>
<td>0.578 (0.584)</td>
<td>0.606 (0.611)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5%</td>
<td>0.040 (0.197)</td>
<td>0.047 (0.134)</td>
<td>0.066 (0.114)</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.079 (0.167)</td>
<td>0.091 (0.137)</td>
<td>0.128 (0.154)</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>0.214 (0.248)</td>
<td>0.239 (0.245)</td>
<td>0.323 (0.333)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5%</td>
<td>-0.021 (0.248)</td>
<td>0.005 (0.134)</td>
<td>0.009 (0.094)</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>0.015 (0.181)</td>
<td>0.041 (0.114)</td>
<td>0.067 (0.109)</td>
<td>10%</td>
</tr>
</tbody>
</table>
Table 3: Infeasible Maximum Likelihood

<table>
<thead>
<tr>
<th>$p_{11} = p_{00}$</th>
<th>n</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References


