Nonparametric Identification of a Binary Random Factor in Cross Section Data

Yingying Dong and Arthur Lewbel[*]

California State University Fullerton and Boston College

Original January 2009, revised March 2011

Abstract

Suppose V and U are two independent mean zero random variables, where V has an asymmetric distribution with two mass points and U has some zero odd moments (having a symmetric distribution suffices). We show that the distributions of V and U are nonparametrically identified just from observing the sum $V + U$, and provide a pointwise rate root n estimator. This can permit point identification of average treatment effects when the econometrician does not observe who was treated. We extend our results to include covariates X, showing that we can nonparametrically identify and estimate cross section regression models of the form $Y = g(X, D^*) + U$, where D^* is an unobserved binary regressor.

JEL Codes: C25, C21
Keywords: Mixture model, Random effects, Binary, Unobserved factor, Unobserved regressor, Nonparametric identification, Deconvolution, Treatment

1 Introduction

We propose a method of nonparametrically identifying and estimating cross section regression models that contain an unobserved binary regressor or treatment, or equivalently an unobserved random effect that can take on two values. For example, suppose an experiment (natural or otherwise) with random or exogenous assignment to treatment was performed on some population, but we only have survey data collected in the

[*] Corresponding Author: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu,http://www2.bc.edu/~lewbel/

region where the experiment occured, and this survey does not report which (or even how many) individuals were treated. Then, given our assumptions, we can point identify the average treatment effect in this population and the probability of treatment, despite not observing who was treated.

No instruments or proxies for the unobserved binary regressor or treatment need to be observed. Identification is obtained by assuming that the unobserved exogenously assigned treatment or binary regressor effect is a location shift of the observed outcome, and that the regression or conditional outcome errors have zero low order odd moments (a sufficient condition for which is symmetrically distributed errors). These identifying assumptions provide moment conditions that can be used to construct either an ordinary generalized method of moments (GMM) estimator, or in the presence of covariates, a nonparametric local GMM estimator for the model.

The zero low order odd moments used for identification here can arise in a number of contexts. Normal errors are of course symmetric and so have all odd moments equal to zero, and normality arises in many models such as those involving central limit theorems, e.g., Gibrat's law for wage or income distributions. Differences of independently, identically distributed errors, or more generally of exchangable errors such as those following ARMA processes, are also symmetrically distributed (see, e.g., proposition 1 of Honore 1992). So, e.g., two period panel models with fixed effects and ARMA errors will generally have errors that are symmetric after time differencing. Our results could therefore be applied in a two period panel where individuals can have an unobserved mean shift at any time (corresponding to the unobserved binary regressor), fixed effects (which are differenced away) and exchangable remaining errors (which yield symmetric errors after differencing). Below we give other more specific examples of models with the required odd moments being zero.

Ignoring covariates for the moment, suppose $Y = h + V + U$, where V and U are independent mean zero random variables and h is a constant. The random V equals either b_0 or b_1 with unknown probabilities p and $1-p$ respectively, where p does not equal a half, i.e., V is asymmetrically distributed. U is assumed to have its first few odd moments equal to zero. We observe a sample of observations of the random variable Y, and so can identify the marginal distribution of Y, but we do not observe h, V, or U.

We first show that the constant h and the distributions of V and U are nonparametrically identified just from observing Y. The only regularity assumption required is that some higher moments of Y exist. More precisely, the first three odd moments of U must be zero (and so also exist for Y) for local identification, while having the first five odd moments of U equal zero suffices for global identification.

We also provide estimators for the distributions of V and U. We show that the constant h, the probability mass function of V, moments of the distribution of U, and points of the distribution function of U can all be estimated using GMM. Unlike common deconvolution estimators that can converge at slow rates, we estimate the distributions of V and U, and the density of U (if it is continuous) at the same rates of convergence as if V and U were separately observed, instead of just observing their sum.

We do not assume that the supports of V or U are known, so estimation of the distribution of V means identifying and estimating both of its support points b_0 and b_1, as well as the probabilities p and $1-p$, respectively, of V equaling b_0 or b_1.

One can write V as $V = b_1 D^* + b_0 (1 - D^*)$ where D^* is an unobserved binary indicator. For example, if D^* is the unobserved indicator of exogenously assigned treatment, then $b_1 - b_0$ is the average treatment effect, p is the probability of treatment, and U describes the remaining heterogeneity of outcomes Y (here treatment is assumed to only cause a shift in outcome means)

We also show how these results can be extended to allow for covariates. If h depends on a vector of covariates X while V and U are independent of X, then we obtain the random effects regression model $Y = h(X) + V + U$, which is popular for panel data, but which we identify and estimate just from cross section data.

More generally, we allow both h and the distributions of V and U to depend in unknown ways on X. This is equivalent to nonparametric identification and estimation of a regression model containing an unobserved binary regressor. The regression model is $Y = g(X, D^*) + U$, where g is an unknown function, D^* is an unobserved binary regressor (or unobserved indicator of treatment) that equals zero with unknown probability $p(X)$ and one with probability $1 - p(X)$, while U is a random error with an unknown conditional distribution $F_U(U \mid X)$ having its first few odd moments equal to zero (conditional symmetry conditioning on X suffices). The unobserved random variables U and D^* are assumed to be conditionally independent, conditioning upon X. By defining $h(x) = E(Y \mid X = x) = E[g(X, D^*) \mid X = x]$, $V = g(X, D^*) - h(X)$ and $U = Y - h(X) - V$, this regression model can then be rewritten as $Y = h(X) + V + U$, where $h(x)$ is a nonparametric regression function of Y on X, and the two support points of V conditional on $X = x$ are then $b_d(x) = g(x, d) - h(x)$ for $d = 0, 1$. Kitamura (2004) provides some nonparametric identification results for this model by placing constraints on how the distributions can depend upon X, while we place no such restrictions on the distribution of X and instead restrict the shape of the distribution of V.

The assumptions we impose on U in $Y = g(X, D^*) + U$ are common assumptions in regression models, e.g., they allow the error U to be heteroskedastic with respect to X, and they hold, e.g., if U given X is normal (though normality is not required). Also, regression model errors U are sometimes interpreted as measurement error in Y, and measurement errors are often assumed to be symmetric.

If D^* is an unobserved treatment indicator, then $g(X, 1) - g(X, 0)$ is the conditional average treatment effect, which may be averaged over X to obtain an unconditional average treatment effect. Symmetry of errors is not usually assumed for treatment models, but suppose we have panel data (two periods of observations) and all treatments occur in one of the two periods. Then as noted above the required symmetry of U errors would result automatically from time differencing the data, given the standard panel model assumption of individual specific fixed effects plus independently, identically distributed (or more generally ARMA or other exchangable) errors.

Another possible application of these extensions is a stochastic frontier model, where Y is the log of a firm's output, X are factors of production, and D^* indicates whether the firm operates efficiently at the frontier, or inefficiently. Existing stochastic frontier models obtain identification either by assuming parametric functional forms for both the distributions of V and U, or by using panel data and assuming that each firm's individual efficiency level is a fixed effect that is constant over time. See, e.g., Kumbhakar et. al. (2007) and Simar and Wilson (2007). In contrast, our assumptions and associated estimators could be used to estimate a nonparametric stochastic frontier model using cross section data, given the restriction that unobserved efficiency is indexed by a binary D^*. Note that virtually all stochastic frontier models based on cross section data assume U given X is symmetrically distributed.

Dong (2008) empirically estimates a model where $Y = h(X) + V + U$, based on symmetry of U and using moments similar to the exponentials we suggest in an extension section. Our results formally prove identification of Dong's model, and our estimator is more general in that it allows V and the distribution of U to depend in arbitrary ways on X. Hu and Lewbel (2007) also nonparametrically identify some features of a model containing an unobserved binary regressor, using either a type of instrumental variable or an assumption of conditional independence of low order moments.

Models that allocate individuals into various types, as D^* does, are common in the statistics and marketing literatures. Examples include cluster analysis, latent class analysis, and mixture models (see, e.g., Clogg 1995 and Hagenaars and McCutcheon

2002). Our model resembles a finite (two distribution) mixture model, but differs crucially in that, for identification, finite mixture models usually require the distributions being mixed to be parametrically specified, while in our model U is nonparametric. While general mixture models are more flexible than ours in allowing for more than two groups and permitting the U distribution to vary across groups, ours is more flexible in letting U be nonparametric, essentially allowing for an infinite number of parameters versus finitely parameterized mixtures.

Some mixture models can be nonparametrically identified by observing draws of vectors of data, where the number of elements of the observed vectors is larger than the number of distributions being mixed. Examples include Hall and Zhou (2003) and Kasahara and Shimotsu (2009). In contrast, we obtain identification with a scalar Y. As noted above, Kitamura (2004) also obtains nonparametric identification with a scalar Y, but does so by requiring observation of a covariate that affects the component distributions with some restrictions. Another closely related mixture model result is Bordes, Mottelet, and Vandekerkhove (2006), who impose strictly stronger conditions than we do, including that U is symmetric and continuously distributed.

Also related is the literature on mismeasured binary regressors, where identification generally requires instruments. An exception is Chen, Hu and Lewbel (2008). Like our Theorem 1 below, they exploit error symmetry for identification, but unlike this paper they assume that the binary regressor is observed, though with some measurement (classification) error, instead of being completely unobserved. A more closely related result is Heckman and Robb (1985), who like us use zero low order odd moments to identify a binary effect, though theirs is a restricted effect that is strictly nested in our results. Error symmetry has also been used to obtain identification in a variety of other econometric contexts, e.g., Powell (1986).

There are a few common ways of identifying the distributions of random variables given just their sum. One method of identification assumes that the exact distribution of one of the two errors is known a priori, (e.g., from a validation sample as is common in the statistics literature on measurement error; see, e.g., Carroll, et. al. 2006) and using deconvolution to obtain the distribution of the other one. For example, if U were normal, one would need to know a priori its mean and variance to estimate the distribution of V. A second standard way to obtain identification is to parameterize both the distributions of V and U, as in most of the latent class literature or in the stochastic frontier literature (see, e.g., Kumbhakar and Lovell 2000) where a typical parameterization is to have V be log normal and U be normal. Panel data models often have errors of the form $V + U$ that are identified either by imposing specific error structures or assuming one of the errors is fixed over time (see, e.g., Baltagi 2008 for a survey of random effects and fixed effects panel data models). Past nonparametric stochastic frontier models have similarly required panel data for identification, as described above. In contrast to all these identification methods, in our model both U and V have unknown distributions, and no panel data are required.

The next section contains our main identification result. We then provide moment conditions for estimating the model, including the distribution of V (its support points and the associated probability mass function), using ordinary GMM. Next we give estimators for the distribution and density function of U. We provide a Monte Carlo analysis showing that our estimator performs reasonably well even compared to infeasible maximum likelihood estimation. This is followed by some extensions showing how our identification and estimation methods can be augmented to provide additional moments for estimation, and to allow for covariates. Proofs are in the Appendix.

2 Identification

In this section, we provide our general identification result. Later we extend these results to include covariates X.

ASSUMPTION A1: Let $Y = h + V + U$. Assume the distribution of V is mean zero, asymmetric, and has exactly two points of support. Assume $E\left(U^d \mid V\right) = E\left(U^d\right)$ exists for all positive integers $d \leq 9$, and $E\left(U^{2d-1}\right) = 0$ for all positive integers $d \leq 5$.

THEOREM 1: Let Assumption A1 hold, and assume the distribution of Y is identified. Then the constant h and the distributions of V and U are identified.

Let b_0 and b_1 denote the two support points of the distribution of V, where without loss of generality $b_0 < b_1$, and let p be the probability that $V = b_0$, so $1 - p$ is the probability that $V = b_1$. Identification of the distribution of V by theorem 1 means identification of b_0, b_1, and p. If Y is the outcome of a treatment and D^* denotes an unobserved treatment indicator, then we can define $V = b_1 D^* + b_0 \left(1 - D^*\right)$, and we will have identification of the probability of treatment p and identification of the average treatment effect $b_1 - b_0$ (which by mean independence of V and U will also equal the average treatment effect on the treated).

Assumption A1 says that the first nine moments of U conditional on V are the same as the moments that would arise if U were distributed symmetrically and independent of V. The only regularity condition required for identification of the distribution of V in Theorem 1 is existence of $E\left(Y^9\right)$.

The proof of Theorem 1 also shows identification up to a finite set, and hence local identification of the distribution of V, assuming just $E(U) = E\left(U^3\right) = E\left(U^5\right) = 0$ and $E\left(U^d \mid V\right) = E\left(U^d\right)$ for positive integers $d \leq 5$, thereby only needing existence of $E\left(Y^5\right)$. Higher moments going up to the ninth moment are required only to distinguish amongst the elements in the finite identified set and thereby provide global identification.

Note that Theorem 1 assumes asymmetry of V (since otherwise it would be indistinguishable from U) and more than one point of support (since otherwise it would be indistinguishable from h), which is equivalent to requiring that p not be exactly equal to zero, one, or one half. This suggests that the identification and associated estimation will be weak if the actual p is very close to any of these values. In practice, it would be easy to tell if this problem exists, because if it does then the observed Y will itself be close to symmetrically distributed. Applying a formal test of data symmetry such as Ahmed and Li (1997) to the Y data is equivalent in our model to testing if p equals zero, one, or one half. More simply, one might look for substantial asymmetry in a histogram or kernel density estimate of Y.

We next consider estimation of h, b_0, b_1, and p, and then later show how the rest of the model, i.e., the distribution function of U, can be estimated.

3 Estimation

Our estimator will take the form of the standard Generalized Method of Moments (GMM, as in Hansen 1982), since given data Y_1,...Y_n, we will below construct a set of moments of the form $E\left[G\left(Y,\theta\right)\right]=0$, where G is a set of known functions and the vector θ consists of the parameters interest h, b_0, p, as well as u_2, u_4, and u_6, where $u_d=E\left(U^d\right)$. The parameters u_2, u_4, and u_6 are nuisance parameters for estimating the V distribution, but in some applications they may be of interest as summary measures of the distribution of U.

Let $v_d=E\left(V^d\right)$. Then $v_1=E\left(V\right)=b_0p+b_1\left(1-p\right)=0$, so

$$b_1=b_0p/\left(p-1\right), \tag{1}$$

and therefore,

$$v_d = E\left(V^d\right) = b_0^d p + \left(\frac{b_0 p}{p-1}\right)^d (1-p)\,. \tag{2}$$

Now expand the expression $E\left[(Y-h)^d - (V+U)^d\right] = 0$ for integers d, noting by Assumption A1 that the first five odd moments of U are zero. The results are

$$E\left(Y-h\right) = 0 \tag{3}$$

$$E\left((Y-h)^2 - (v_2 + u_2)\right) = 0 \tag{4}$$

$$E\left((Y-h)^3 - v_3\right) = 0 \tag{5}$$

$$E\left((Y-h)^4 - (v_4 + 6v_2u_2 + u_4)\right) = 0 \tag{6}$$

$$E\left((Y-h)^5 - (v_5 + 10v_3u_2)\right) = 0 \tag{7}$$

$$E\left((Y-h)^6 - (v_6 + 15v_4u_2 + 15v_2u_4 + u_6)\right) = 0 \tag{8}$$

$$E\left((Y-h)^7 - (v_7 + 21v_5u_2 + 35v_3u_4)\right) = 0 \tag{9}$$

$$E\left((Y-h)^9 - (v_9 + 36v_7u_2 + 126v_5u_4 + 84v_3u_6)\right) = 0 \tag{10}$$

Substituting equation (2) into equations (3) to (10) gives eight moments we can write as $E\left[G\left(Y,\theta\right)\right] = 0$ in the six unknown parameters $\theta = (h,\ b_0,\ p,\ u_2,\ u_4,\ u_6)$, which we use for estimation via GMM. We use these moments because the proof of Theorem 1 shows that these particular eight equations are exactly those required to point identify the parameters defining the distribution of V. As shown in the proof, more equations than unknowns are required for global identification because of the nonlinearity of these equations, and in particular the presence of multiple roots. Given an estimate of θ, the estimate of b_1 is then obtained by equation (1).

Based on Theorem 1, for this estimator we assume that Y_1, ...,Y_n are identically distributed (or more precisely, have identical first nine moments), however, the Y observations do not need to be independent, since GMM estimation theory permits some serial dependence in the data. Standard GMM limiting distribution theory applied to our moments provides root n consistent, asymptotically normal estimates of θ and hence of h and of the distribution of V, (i.e., the support points b_0 and b_1 and the probability p, where $\widehat{b}_1$ is obtained by $\widehat{b}_1 = \widehat{b}_0\widehat{p}/\left(\widehat{p}-1\right)$ from equation 1). Without loss of generality we have imposed $b_0 < b_1$ (if this is violated then the definitions of these two parameters can be switched to make the inequality hold), and this along with $E\left(V\right) = 0$ implies that $\widehat{b}_0$ is negative and $\widehat{b}_1$ is positive, which may be imposed on estimation. One could also impose that $\widehat{p}$ lie between zero and one, and that $\widehat{u}_2$, $\widehat{u}_4$, and $\widehat{u}_6$ be positive.

One might anticipate poor empirical results, and great sensitivity to outliers or extreme observations of Y, given the use of such high order moments for estimation. For example, Altonji and Segal (1996) document bias in GMM estimates associated with just second order moments. However, we found that these problems rarely arose in our monte carlo simulations (in applications one would want to carefully scale Y to avoid the effects of computer rounding errors based on inverting matrix entries of varying orders of magnitude). We believe the reason the estimator performs reasonably well is that only lower order moments are required for local identification, up to a small finite set of values. The higher order moments (specifically those above the fifth) are only needed for global identification to distinguish between these few possible multiple solutions of the low order polynomials. For example, if by the low order moments b_1 was identified up to a value in the neighborhood of either 1 or of 3, then even poorly estimated higher moments could succeed in distinguishing between these two neighborhoods, by having a sample GMM moment mean be substantially closer to zero in the neighborhood of one value rather than the other.

In an extension section we describe how additional moments could be constructed for estimation based on symmetry of U. These alternative moments might be employed in applications where the polynomial based moments are found to be problematic. Another possibility would be to Winsorize extreme observations of the Y data prior to estimation to robustify the higher moment estimates.

4 The Distribution of U

For any random variable Z, let F_Z denote the marginal cumulative distribution function of Z. Define $\varepsilon = V + U$. Define $F_{KU}(u)$ by

$$\begin{aligned} F_{KU}(u) &= \sum_{k=0}^{K-1} \left(\frac{1-p}{p}\right)^k \frac{1}{p} F_\varepsilon\left(u + b_0 + (b_0 - b_1)k\right) \quad \text{if } p > 1/2 \quad \text{otherwise} \\ F_{KU}(u) &= \sum_{k=0}^{K-1} \left(\frac{p}{1-p}\right)^k \frac{1}{1-p} F_\varepsilon\left(u + b_1 + (b_1 - b_0)k\right) \quad \text{if } p < 1/2 \end{aligned}$$

The proof of Theorem 1 shows that $F_U(u) = F_{KU}(u) + R_K$, where $0 \leq R_K \leq \min\left(\left(\frac{p}{1-p}\right)^K, \left(\frac{1-p}{p}\right)^K\right)$, so the remainder term $R_K \to 0$ as $K \to \infty$ (since $p = 1/2$ is ruled out). This suggests that F_U could be estimated by $F_{KU}(u)$ after replacing F_ε with the empirical distribution of $Y - \widehat{h}$, replacing b_0, b_1, and p with their estimates, and letting $K \to \infty$ as $N \to \infty$.

However, under the assumption that U is symmetrically distributed, the following theorem provides a more convenient way to estimate the distribution function of U. Define

$$\Psi(u) = \frac{\left[F_\varepsilon(-u + b_0) - 1\right]p + F_\varepsilon(u + b_1)(1-p)}{1 - 2p}. \tag{11}$$

THEOREM 2: Let Assumption A1 hold. Assume U is symmetrically distributed. Then

$$F_U(u) = \frac{\Psi(u) - \Psi(-u) + 1}{2}. \tag{12}$$

Theorem 2 provides a direct expression for the distribution of U in terms of b_0, b_1,

p and the distribution of ε, all of which are previously identified. This can be used to construct an estimator for $F_U(u)$ as follows.

Let $I(\cdot)$ denote the indicator function that equals one if $\cdot$ is true and zero otherwise, and let θ be a vector containing h, b_0, b_1, and p. Define the function $\omega(Y, u, \theta)$ by

$$\omega(Y, u, \theta) = \frac{\left[I(Y \leq h - u + b_0) - 1\right]p + I(Y \leq h + u + b_1)(1 - p)}{1 - 2p}. \tag{13}$$

Then using $Y = h + \varepsilon$ it follows immediately from equation (11) that

$$\Psi(u) = E(\omega(Y, u, \theta)). \tag{14}$$

An estimator for $F_U(u)$ can now be constructed by replacing the parameters in equation (14) with estimates, replacing the expectation with a sample average, and plugging the result into equation (12). The resulting estimator is

$$\widehat{F}_U(u) = \frac{1}{n}\sum_{i=1}^{n} \frac{\omega\left(Y_i, u, \widehat{\theta}\right) - \omega\left(Y_i, -u, \widehat{\theta}\right) + 1}{2}. \tag{15}$$

Alternatively, $F_U(u)$ for a finite number of values of u, say $u_1, ..., u_J$, can be estimated as follows. Recall that $E[G(Y, \theta)] = 0$ was used to estimate the parameters h, b_0, b_1, p by GMM. For notational convenience, let $\eta_j = F_U(u_j)$ for each u_j. Then by equations (12) and (14),

$$E\left[\eta_j - \frac{\omega(Y, u_j, \theta) - \omega(Y, u_j, \theta) + 1}{2}\right] = 0. \tag{16}$$

Adding equation (16) for $j = 1, ..., J$ to the set of functions defining G, including $\eta_1, ..., \eta_J$ in the vector θ, and then applying GMM to this augmented set of moment conditions $E[G(Y, \theta)] = 0$ simultaneously yields root n consistent, asymptotically normal estimates of h, b_0, b_1, p and $\eta_j = F_U(u_j)$ for $j = 1, ..., J$. An advantage of this approach versus equation (15) is that GMM limiting distribution theory then provides standard error estimates for each $\widehat{F}_U(u_j)$.

While p is the unconditional probability that $V = b_0$, given $\widehat{F}_U$ it is straightforward to estimate conditional probabilities as well. In particular,

$$\begin{aligned}
\Pr(V = b_0 \mid Y \leq y) &= \Pr(V = b_0, Y \leq y) / \Pr(Y \leq y) \\
&= F_U(y - h - b_0) / F_y(y)
\end{aligned}$$

which could be estimated as $\widehat{F}_U\left(y - \widehat{h} - \widehat{b}_0\right) / \widehat{F}_y(y)$ where $\widehat{F}_y$ is the empirical distribution of Y.

Let f_Z denote the probability density function of any continuously distributed random variable Z. So far no assumption has been made about whether U is continuous or discrete. However, if U is continuous, then ε and Y are also continuous, and then taking the derivative of equations (11) and (12) with respect to u gives

$$\psi(u) = \frac{-f_\varepsilon(-u + b_0)\,p + f_\varepsilon(u + b_1)(1 - p)}{1 - 2p}, \qquad f_U(u) = \frac{\psi(u) + \psi(-u)}{2}, \tag{17}$$

which suggests the estimators

$$\widehat{\psi}(u) = \frac{-\widehat{f}_\varepsilon\left(-u + \widehat{b}_0\right)\widehat{p} + \widehat{f}_\varepsilon\left(u + \widehat{b}_1\right)(1 - \widehat{p})}{1 - 2\widehat{p}}, \tag{18}$$

$$\widehat{f}_U(u) = \frac{\widehat{\psi}(u) + \widehat{\psi}(-u)}{2}, \tag{19}$$

where $\widehat{f}_\varepsilon(\varepsilon)$ is a kernel density or other estimator of $f_\varepsilon(\varepsilon)$, constructed using data $\widehat{\varepsilon}_i = Y_i - \widehat{h}$ for $i = 1, ...n$. Since densities converge at slower than rate root n, the limiting distribution of this estimator will generally be the same as if $\widehat{h}$, $\widehat{b}_0$, $\widehat{b}_1$, and $\widehat{p}$

were evaluated at their true values (e.g., this holds if f is differentiable by a mean value expansion of $\widehat{\psi}$ around the true values of h, b_0, b_1, and p). The above $\widehat{f_U}(u)$ is just the weighted sum of two density estimators, each one dimensional, and so will converge at the same rate as a one dimensional density estimator. For example, this will be the pointwise rate $n^{-2/5}$ using a kernel density estimator of $\widehat{f_\varepsilon}$ under standard assumptions as in Silverman (1986) (independent observations, f_ε twice differentiable, evaluated at points not on the boundary of the support of ε, bandwith proportional to $n^{-1/5}$, and a second order kernel function) with p bounded away from 1/2. It is possible for $\widehat{f_U}(u)$ to be negative in finite samples, so if desired one could replace negative values of $\widehat{f_U}(u)$ with zero.

A potential numerical problem is that equation (18) may require evaluting $\widehat{f_\varepsilon}$ at a value that is outside the range of observed values of $\widehat{\varepsilon}_i$. Since both $\widehat{\psi}(u)$ and $\widehat{\psi}(-u)$ are consistent estimators of $\widehat{f_U}(u)$ (though generally less precise than equation (19) because they individually ignore the symmetry constraint), one could use either $\widehat{\psi}(u)$ or $\widehat{\psi}(-u)$ instead of their average to estimate $\widehat{f_U}(u)$ whenever $\widehat{\psi}(-u)$ or $\widehat{\psi}(u)$, respectively, requires evaluating $\widehat{f_\varepsilon}$ at a point outside the range of observed values of $\widehat{\varepsilon}_i$.

This construction also suggests a specification test for the model. Since symmetry of U implies that $\widehat{\psi}(u) = \widehat{\psi}(-u)$ one could base a test on whether $\int_0^L \left[\widehat{\psi}(u) - \widehat{\psi}(-u)\right]^2 w(u)\,du = 0$, where $w(u)$ is a weighting function that integrates to one, and L is in the range of values for which neither $\widehat{\psi}(-u)$ nor $\widehat{\psi}(u)$ requires evaluating $\widehat{f_\varepsilon}$ at a point outside the range of observed values of $\widehat{\varepsilon}_i$. The limiting distribution theory for this type of test statistic (a degenerate U statistic under the null) based on functions of kernel densities is standard, and in this case would closely resemble Ahmed and Li (1997).

5 Monte Carlo Analysis

Our Monte Carlo design takes $h = 0$, U standard normal, and $-b_0 p = b_1 (1 - p) = 1$. We consider three different values of p between one half and one, specifically, .6, .8, and .95. By symmetry of this design, we should obtain the same results in terms of accuracy if we took p equal to .4, .2, and .05, respectively. The nuisance parameters, derived from the distribution of U are then $u_2 = 1$, $u_4 = 3$, and $u_6 = 15$. Our sample size is $n = 1000$, and for each design we perform 1000 Monte Carlo replications. Each draw of Y in each replication is constructed by drawing an observation of V and one of U from their above described distributions and then summing the two.

In each simulated data set we first estimated h as the sample mean of Y, then performed standard two step GMM (using the identity matrix as the weighting matrix in the first step) with the moment equations (4) to (10). We imposed the inequality constraints on estimation that p lie between zero and one and that b_1, u_2, u_4, and u_6 are positive. Rarely, this GMM either failed to converge after many iterations, or iterated towards one of these boundary points. When this happened, we applied two step GMM to just the low order (sufficient for local identification) moments (4), (5), (7), and then used the results as starting values for two step GMM using all the moments (4) to (10). We could have alternatively performed a more time consuming grid search, but this procedure led to estimates that conveged to interior points in all but a handful of simulations. Specifically, in fewer than one half of one percent of replications this procedure either failed to converge or produced estimates of p that approached the boundaries of zero or one. We drop these few failed replications from our reported results.

The results are reported in Tables 1, 2, and 3. The parameters of the distribution of V (b_0, b_1, and p) are estimated with reasonable accuracy, having relatively small root mean squared errors and interquartile ranges. These parameters are very close to median unbiased, but have mean bias of a few percent, with p always mean biased downwards. This is likely because estimates of $\widehat{p}$ are more or less equally likely to be above or below the true (yielding very small median bias), but when they are below the true they can be much further from the truth than when they are too high, e.g., when $p = .8$ an estimate that is too high can be biased by at most .2, while the downward bias can be as large as $-.8$. Some fraction of these replications may be centering around incorrect roots of the polynomial moments, which can be quite distant from the correct roots.

The high order nuisance parameters U_4 and U_6 are sometimes estimated very poorly. In particular, for $p = .6$ the median bias of U_6 is almost -20% while the mean bias is over five times larger and has the opposite sign of the median bias. The fact that the high order moment nuisance parameters are generally much more poorly estimated than the parameters of interest supports our claim that low order moments are providing most of the parameter estimation precision, while higher order moments mainly serve to distinguish among discretely separated local alternatives.

We performed limited experiments with alternative sample sizes, which are not reported to save space. Precision increases with sample size pretty much as one would expect. More substantial is the frequency with which numerical problems were encountered, e.g., at $n = 500$ we encountered convergence or boundary problems in 1.4% of replications while these problems were almost nonexistent at $n = 5000$.

6 Extension 1: Additional Moments

Here we provide additional moments that might be used for estimating the parameters h, b_0, b_1 and p.

PROPOSITION 1: Let $Y = h + V + U$. Assume the distribution of V is mean zero, asymmetric, and has exactly two points of support. Assume U is symmetrically distributed around zero and is independent of V. Assume $E\left[\exp\left(TU\right)\right]$ exists for some positive constant T. Then for any positive $\tau \leq T$ there exists a constant α_τ such that the following two equations hold with $r = p/\left(1-p\right)$:

$$E\left[\exp\left(\tau\left(Y-h\right)\right) - \left(r\exp\left(\tau b_0\right) + \exp\left(-\tau r\right)\right)\alpha_\tau\right] = 0 \tag{20}$$

$$E\left[\exp\left(-\tau\left(Y-h\right)\right) - \left(r\exp\left(-\tau b_0\right) + \exp\left(\tau b_0\right)\right)\alpha_\tau\right] = 0 \tag{21}$$

Given a set of L positive values for τ, i.e., constants τ_1,...,τ_L, each of which are less than T, equations (20) and (21) provide $2L$ moment conditions satisfied by the set of $L+3$ parameters α_{τ_1},..., α_{τ_L}, h, p, and b_0. Although the order condition for identification is therefore satisfied with $L \geq 3$, we do not have a proof analogous to Theorem 1 showing that the parameters are actually globally identified based on any number of these moments. Also, Proposition 1 is based on means of exponents, and so requires Y to have a thinner tailed distribution than estimation based on the polynomial equations (3) to (10). Still, if global identification holds with these parameters, then they could be used by themselves for estimation, otherwise they could be combined with the polynomial moments to possibly increase estimation efficiency.

One could also construct moments of complex exponentials based on the characteristic function of $Y-h$ instead of those based on the moment generating function as in Proposition 1, which avoids the requirement for thin tailed distributions. However such moments could sometimes vanish and thereby be uninformative, as when U is uniform.

Proposition 1 actually provides a continuum of moments, so rather than just choose a finite number of values for τ, it would also be possible to efficiently combine all the moments given by an interval of values of τ using, e.g., Carrasco and Florens (2000).

7 Extension 2: h depends on covariates

We now extend our results by permitting h to depend on covariates X. Estimators associated with this extension will take the form of standard two step estimators with a uniformly consistent first step.

COROLLARY 1: Assume the conditional distribution of Y given X is identified and its mean exists. Let $Y = h(X) + V + U$. Let Assumption A1 hold. Assume V and U are independent of X. Then the function $h(X)$ and distributions of U and V are identified.

Corollary 1 extends Theorem 1 by allowing the conditional mean of Y to nonparametrically depend on X. Given the assumptions of Corollary 1, it follows immediately that equations (3) to (10) hold replacing h with $h(X)$, and if U is symmetrically distributed and independent of V and X then equations (20) and (21) also hold replacing h with $h(X)$. This suggests a couple of ways of extending the GMM estimators of the previous section. One method is to first estimate $h(X)$ by a uniformly consistent nonparametric mean regression of Y on X (e.g., a kernel regression over a compact set of X values on the interior of its support), then replace $Y - h$ in equations (3) to (10) and/or equations (20) and (21) with $\varepsilon = Y - h(X)$, and apply ordinary GMM to the resulting moment conditions (using as data $\widehat{\varepsilon}_i = Y_i - \widehat{h}(X_i)$ for $i = 1, ..., n$) to estimate the parameters b_0, b_1, p, u_2, u_4, and u_6. Consistency of this estimator follows immediately from the uniform consistency of $\widehat{h}$ and ordinary consistency of GMM. This estimator is easy to implement because it only depends on ordinary nonparametric regression and ordinary GMM. Root n limiting distribution theory may be immediately obtained by applying generic two step estimation theorems as in Newey and McFadden (1994).

After replacing $\widehat{h}$ with $\widehat{h}(X_i)$, equation (15) can be used to estimate the distribution of U, or alternatively equation (16) for $j = 1, ..., J$, replacing h with $h(X)$, can be included in the set of functions defining G in the estimator described above. Since ε has the same properties here as before, given uniform consistency of $\widehat{h}(X)$, the estimator (19) will still consistently estimate the density of U if it is continuous, using as data $\widehat{\varepsilon}_i = Y_i - \widehat{h}(X_i)$ for $i = 1, ..., n$ to estimate the density function f_ε.

8 Extension 3: Nonparametric regression with an Unobserved Binary Regressor

This section extends previous results to a more general nonparametric regression model of the form $Y = g(X, D^*) + U$. Specifically, we have the following corollary.

COROLLARY 2: Assume the joint distribution of Y, X is identified and that $g(X, D^*) = E(Y \mid X, D^*)$ exists, where D^* is an unobserved variable with support $\{0, 1\}$. Assume that the distribution of $g(X, D^*)$ conditional upon X is asymmetric for all X on its support. Define $p(X) = E(1 - D^* \mid X)$ and define $U = Y - g(X, D^*)$. Assume $E\left(U^d \mid X, D^*\right) = E\left(U^d \mid X\right)$ exists for all integers $d \leq 9$ and $E\left(U^{2d-1} \mid X\right) = 0$ for all positive integers $d \leq 5$. Then the functions $g(X, D^*)$, $p(X)$, and the distribution of U are identified.

Corollary 2 permits all of the parameters of the model to vary nonparametrically with X. It provides identification of the regression model $Y = g(X, D^*) + U$, allowing the unobserved model error U to be heteroskedastic (and have nonconstant higher moments as well), though the variance and other low order even moments of U can only depend on X and not on the unobserved regressor D^*. As noted in the introduction and in the proof of this Corollary, $Y = g(X, D^*) + U$ is equivalent to $Y = h(X) + V + U$. However, unlike Corollary 1, now V and U have distributions that can depend on X. As with Theorem 1, symmetry of U (now conditional on X) suffices to make the required low order odd moments of U be zero.

Given the assumptions of Corollary 2, equations (3) to (10), and given symmetry of U, equations (20) and (21), will all hold after replacing the parameters h, b_0, b_1, p, u_j, and τ_ℓ, and with functions $h(X)$, $b_0(X)$, $b_1(X)$, $p(X)$, $u_j(X)$, and $\tau_\ell(X)$ and replacing the unconditional expectations in these equations with conditional expectations, conditioning on $X = x$. If desired, we can further replace $b_0(X)$ and $b_1(X)$ with $g(x,0) - h(x)$ and $g(x,1) - h(x)$, respectively, to directly obtain estimates of the function $g(X, D^*)$ instead of $b_0(X)$ and $b_1(X)$.

Let $q(x)$ be the vector of all of the above listed unknown functions. Then these conditional expectations can be written as

$$E[G(q(x), Y) \mid X = x)] = 0 \tag{22}$$

for a vector of known functions G. Equation (22) is in the form of conditional GMM which could be estimated using Ai and Chen (2003), replacing all of the unknown functions $q(x)$ with sieves (related estimators are Carrasco and Florens 2000 and Newey and Powell 2003). However, given independent, identically distributed draws of X, Y, the local GMM estimator of Lewbel (2007) may be easier to use because it exploits the special structure we have here where all the functions $q(x)$ to be estimated depend on the same variables that the moments are conditioned upon, that is, $X = x$. We summarize here how this local GMM estimator would be implemented. See the online supplement to this paper or Lewbel (2007) for details regarding the associated limiting distribution theory.

1. For any value of x, construct data $Z_i = K((x - X_i)/b)$ for $i = 1, ..., n$, where K is an ordinary kernel function (e.g., the standard normal density function) and b is a bandwidth parameter. As is common practice when using kernel functions, it is a good

idea to first standardize the data by scaling each continuous element of X by its sample standard deviation.

2. Obtain $\widehat{\theta}$ by applying standard two step GMM based on the moment conditions $E\left(G\left(\theta, Y\right) Z\right) = 0$ for G from equation (22).

3. For the given value of x, let $\widehat{q}(x) = \widehat{\theta}$.

4. Repeat these steps using every value of x for which one wishes to estimate the vector of functions $q(x)$. For example, one may repeat these steps for a fine grid of x points on the support of X, or repeat these steps for x equal to each data point X_i to just estimate the functions $q(x)$ at the observed data points.

Note that this local GMM estimator can be used when X contains both continuous and discretely distributed elements. If all elements of X are discrete, then the estimator simplifies back to Hansen's (1982) original GMM.

9 Discrete V With More Than Two Support Points

A simple counting argument suggests that it may be possible to extend this paper's identification and associated estimators to applications where V is discrete with more than two points of support, as follows. Suppose V takes on the values b_0, b_1, ..., b_H with probabilities p_0, p_1,..., p_H. Let $u_j = E\left(U^j\right)$ for integers j as before. Then for any positive odd integer S, the moments $E\left(Y^s\right)$ for $s = 1$, ..., S equal known functions of the $2H + \left(S + 1\right)/2$ parameters b_1, b_2,..., b_H, p_1, p_2, ...,p_H, u_2, u_4, ..., u_{S-1}, h. Note p_0 and b_0 can be expressed as functions of the other parameters by probabilities summing to one and V having mean zero, and we assume u_s for odd values of $s \leq S$ are zero. Therefore, with any odd $S \geq 4H + 1$, $E\left(Y^s\right)$ for $s = 1$, ..., S provides at least as many moment equations as unknowns, which could be used to estimate these parameters by GMM and will generally suffice for local identification. These moments include polynomials with up to $S-1$ roots, so having S much larger than $4H+1$ may be necessary for global identification, just as the proof of Theorem 1 requires $S=9$ even though in that theorem $H=1$. Still, as long as U has sufficiently thin tails, $E\left(Y^{s}\right)$ can exist for arbitrarily high integers s, thereby providing far more identifying equations than unknowns.

The above analysis is only suggestive. We do not have a proof of global identification with more than two points of support, though local identification up to a finite set should hold, given that the moments are polynomials, which must have a finite number of roots.

Assuming that a given model where V takes on more than two values is identified, moment conditions for estimation analogous to those we provided earlier are available. For example, as in the proof of Proposition 1 it follows from symmetry of U that

$$E\left[\exp\left(\tau\left(Y-h\right)\right)\right]=E\left[\exp\left(\tau V\right)\right]\alpha_{\tau}$$

with $\alpha_{\tau}=\alpha_{-\tau}$ for any τ for which these expectations exist, and therefore by choosing constants τ_1,...,τ_L, GMM estimation could be based on the $2L$ moments

$$\begin{aligned} E\left[\sum_{k=0}^{H}\left[\left[\exp\left(\tau_{\ell}\left(Y-h\right)\right)\right]-\exp\left(\tau_{\ell}b_k\right)\alpha_{\tau_{\ell}}\right]p_k\right] &= 0 \\ E\left[\sum_{k=0}^{H}\left[\left[\exp\left(-\tau_{\ell}\left(Y-h\right)\right)\right]-\exp\left(-\tau_{\ell}b_k\right)\alpha_{\tau_{\ell}}\right]p_k\right] &= 0 \end{aligned}$$

for $\ell=1,...,L$. The number of parameters b_k, p_k and $\alpha_{\tau_{\ell}}$ to be estimated would be $2H+L$, so taking $L>2H$ provides more moments than unknowns.

10 Conclusions

We have proved global point identification and provided estimators for the models $Y=h+V+U$ or $Y=h(X)+V+U$, and more generally for $Y=g(X,D^{*})+U$. In these models, D^* or V are unobserved regressors with two points of support, and the unobserved U is drawn from an unknown distribution having some odd central moments equal to zero, as would be the case if U is symmetrically distributed. No instruments, measures, or proxies for D^* or V are observed. A small Monte Carlo analysis shows that our estimator works reasonably well with a moderate sample size, despite involving high order data moments.

To further illustrate the estimator, in an online supplemental appendix to this paper we provide a small empirical application involving distribution of income across countries.

Interesting work for the future could include derivation of semiparametric efficiency bounds for the model, and obtaining conditions for global identification when V can take on more than two values.

References

[1] Ai, C. and X. Chen (2003), "Efficient Estimation of Models With Conditional Moment Restrictions Containing Unknown Functions," Econometrica, 71, 1795-1844.

[2] Ahmed, I. A. and Q. Li (1997), "Testing Symmetry of an Unknown Density by Kernel Method," Nonparametric Statistics, 7, 279-293.

[3] Altonji, J. and L. Segal (1994), "Small-Sample Bias in GMM Estimation of Covariance Structures," Journal of Business & Economic Statistics, 14, 353-66.

[4] Baltagi, B. H. (2008), Econometric Analysis of Panel Data, 4th ed., Wiley.

[5] Bordes, L., S. Mottelet and P. Vandekerkhove, (2006) "Semiparametric Estimation of a Two-Component Mixture Model," Annals of Statistics, 34, 1204-1232.

[6] Carrasco, M. and J. P. Florens (2000), "Generalization of GMM to a Continuum of Moment Conditions," Econometric Theory, 16, 797-834.

[7] Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, (2006), Measurement Error in Nonlinear Models: A Modern Perspective, 2nd edition, Chapman & Hall/CRC.

[8] Chen, X., Y. Hu, and A. Lewbel, (2008) "Nonparametric Identification of Regression Models Containing a Misclassified Dichotomous Regressor Without Instruments," Economics Letters, 2008, 100, 381-384.

[9] Chen, X., O. Linton, and I. Van Keilegom, (2003) "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," Econometrica, 71, 1591-1608,

[10] Clogg, C. C. (1995), Latent class models, in G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), Handbook of statistical modeling for the social and behavioral sciences (Ch. 6; pp. 311-359). New York: Plenum.

[11] Dong, Y., (2008), "Nonparametric Binary Random Effects Models: Estimating Two Types of Drinking Behavior," Unpublished manuscript.

[12] Gozalo, P, and Linton, O. (2000). Local Nonlinear Least Squares: Using Parametric Information in Non-parametric Regression. Journal of econometrics, 99, 63-106.

[13] Hagenaars, J. A. and McCutcheon A. L. (2002), Applied Latent Class Analysis Models, Cambridge: Cambridge University Press.

[14] Hall, P., and X.-H. Zhou (2003): "Nonparametric Estimation of Component Distributions in a Multivariate Mixture,"Annals of Statistics, 31, 201–224.

[15] Hansen, L., (1982), "Large Sample Properties of Generalized Method of Moments Estimators," Econometrica, 50, 1029-1054.

[16] Heckman, J. J. and R. Robb, (1985), "Alternative Methods for Evaluating the Impact of Interventions, " in Longitudinal Analysis of Labor Market Data. James J. Heckman and B. Singer, eds. New York: Cambridge University Press, 156-245.

[17] Honore, B. (1992),"Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," Econometrica, 60, 533-565.

[18] Hu, Y. and A. Lewbel, (2008) "Identifying the Returns to Lying When the Truth is Unobserved," Boston College Working paper.

[19] Kasahara, H. and Shimotsu, K. (2009), "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," Econometrica, 77, 135-175.

[20] Kitamura, Y. (2004), "Nonparametric Identifiability of Finite Mixtures," Unpublished Manuscript, Yale University.

[21] Kumbhakar, S. C. and C. A. K. Lovell , (2000), Stochastic Frontier Analysis, Cambridge University Press.

[22] Kumbhakar, S.C., B.U. Park, L Simar, and E.G. Tsionas, (2007) "Nonparametric stochastic frontiers: A local maximum likelihood approach," Journal of Econometrics, 137, 1-27.

[23] Lewbel, A. (2007) "A Local Generalized Method of Moments Estimator," Economics Letters, 94, 124-128.

[24] Lewbel, A. and O. Linton, (2007) "Nonparametric Matching and Efficient Estimators of Homothetically Separable Functions," Econometrica, 75, 1209-1227.

[25] Li, Q. and J. Racine (2003), "Nonparametric estimation of distributions with categorical and continuous data," Journal of Multivariate Analysis, 86, 266-292

[26] Newey, W. K. and D. McFadden (1994), ”Large Sample Estimation and Hypothesis Testing,” in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[27] Newey, W. K. and J. L. Powell, (2003), "Instrumental Variable Estimation of Nonparametric Models," Econometrica, 71 1565-1578.

[28] Powell, J. L. (1986), "Symmetrically Trimmed Least Squares Estimation of Tobit Models," Econometrica, 54, 1435-1460.

[29] Silverman, B. W. (1986), Density Estimation for Statistics and Data Analysis, London: Chapman and Hall.

[30] Simar, L. and P. W. Wilson (2007) "Statistical Inference in Nonparametric Frontier Models: Recent Developments and Perspectives," in The Measurement of Productive Efficiency, 2nd edition, chapter 4, ed. by H. Fried, C.A.K. Lovell, and S.S. Schmidt, Oxford: Oxford University Press.

11 Appendix A: Proofs

PROOF of Theorem 1: To save space, a great deal of tedious but straightforward algebra is omitted. These details are available in an online supplemental appendix.

First identify h by $h = E(Y)$, since V and U are mean zero. Then the distribution of ε defined by $\varepsilon = Y - h$ is identified, and $\varepsilon = U + V$. Define $e_d = E\left(\varepsilon^d\right)$, $u_d = E\left(U^d\right)$, and $v_d = E\left(V^d\right)$. Now evaluate e_d for integers $d \leq 9$. These e_d exist by assumption, and are identified because the distribution of ε is identified. Using independence of V and U, $v_1 = 0$, and $u_d = 0$ for odd values of d up to nine, evaluate $e_d = E\left((U+V)^d\right)$ to obtain $e_2 = v_2 + u_2$, $e_3 = v_3$, $e_4 = v_4 + 6v_2u_2 + u_4$ so $u_4 = e_4 - v_4 - 6v_2e_2 + 6v_2^2$, and $e_5 = v_5 + 10v_3u_2 = v_5 + 10v_3\left(e_2 - v_2\right)$. Define $s = e_5 - 10e_3e_2$, and note that s depends only on identified objects and so is identified. Then $s = v_5 - 10e_3v_2$.

Similarly, $e_6 = v_6 + 15v_4u_2 + 15v_2u_4 + u_6$ which can be solved for u_6, $e_7 = v_7 + 21v_5u_2 + 35v_3u_4$, and $e_9 = v_9 + 36v_7u_2 + 126v_5u_4 + 84v_3u_6$. Substituting out the earlier

expressions for u_2, u_4, and u_6 in the e_7 and e_9 equations gives results that can be written as $q = v_7 - 35e_3v_4 - 21sv_2$ and $w = v_9 - 36qv_2 - 126sv_4 - 84e_3v_6$ where q and w are identified by $q = e_7 - 21se_2 - 35e_3e_4 = e_7 - 21e_5e_2 + e_3\left(210e_2^2 - 35e_4\right)$ and $w = e_9 - 36qe_2 - 126se_4 - 84e_3e_6$
$= e_9 - 36e_7e_2 + e_5\left(756e_2^2 - 126e_4\right) + e_3\left(2520e_2e_4 - 84e_6 - 7560e_2^3\right)$.

Summarizing, we have w, s, q, e_3 are all identified and $e_3 = v_3$, $s = v_5 - 10e_3v_2$, $q = v_7 - 35e_3v_4 - 21sv_2$, and $w = v_9 - 84e_3v_6 - 126sv_4 - 36qv_2$.

Now V only takes on two values, so let V equal b_0 with probability p_0 and b_1 with probability p_1. Let $r = p_0/p_1$. Using $p_1 = 1 - p_0$ and $E\left(V\right) = b_0p_0 + b_1p_1 = 0$ we have

$$p_0 = r/\left(1+r\right), \qquad p_1 = 1/\left(1+r\right), \qquad b_1 = -b_0r,$$

and for any integer d

$$v_d = b_0^dp_0 + b_1^dp_1 = b_0^d\left(p_0 + \left(-r\right)^dp_1\right) = b_0^d\left[r + \left(-r\right)^d\right]/\left(1+r\right).$$

Substituting this v_d into the expression for e_3, s, q, and w reduces to

$$\begin{aligned} e_3 &= b_0^3r\left(1-r\right), \qquad s = b_0^5r\left(1-r\right)\left(r^2 - 10r + 1\right) \\ q &= b_0^7r\left(1-r\right)\left(r^4 - 56r^3 + 246r^2 - 56r + 1\right) \\ w &= b_0^9r\left(1-r\right)\left(r^6 - 246r^5 + 3487r^4 - 10452r^3 + 3487r^2 - 246r + 1\right) \end{aligned}$$

These are four equations in the two unknowns b_0 and r. We require all four equations for point identification, because these are polynomials in r and so have multiple roots. However, from just the e_3 and s equations and $b_0 \neq 0$ we have the identified polynomial in r

$$e_3^5\left(r^2 - 10r + 1\right)^3 - s^3r^2\left(1-r\right)^2 = 0$$

which has at most six roots. Associated with each possible root r is a corresponding identified distribution for V and U as described at the end of this proof. This shows set identification of the model up to a finite set, and hence local identification, using just the first five moments of Y.

To show global identification, we will first show that the four equations for e_3, s, q, and w imply that $r^2 - \gamma r + 1 = 0$, where γ is finite and identified.

First we have $e_3 = v_3 \neq 0$ and $r \neq 1$ by asymmetry of V. Also $r \neq 0$ and $b_0 \neq 0$ because then V would only have one point of support instead of two. Applying these results to the s equation shows that if s (which is identified) is zero then $r^2-10r+1=0$, and so in that case γ is identified. So now consider the case where $s\neq 0$.

Define $R=qe_3/s^2$, which is identified because its components are identified. Then

$$R=\left(r^4-56r^3+246r^2-56r+1\right)\left(r^2-10r+1\right)^{-2} \qquad \text{so}$$

$$0=(1-R)\,r^4+(-56+20R)\,r^3+(246-102R)\,r^2+(-56+20R)\,r+(1-R)$$

If $R=1$, then (using $r\neq 0$) this polynomial reduces to the quadratic $0=r^2-4r+1$, so in this case $\gamma=-4$ is identified. Now consider the case where $R\neq 1$.

Define $Q=s^3/e_3^5$ and $S=w/e_3^3$. Both Q and S exist because $e_3\neq 0$, and they are identified because their components are identified. Then

$$\begin{aligned} Q &= \left(r^2-10r+1\right)^3\left(r\left(1-r\right)\right)^{-2} \quad \text{so} \\ 0 &= r^6-30r^5+(303-Q)\,r^4+(2Q-1060)\,r^3+(303-Q)\,r^2-30r+1 \end{aligned}$$

Also

$$\begin{aligned} \frac{w}{e_3^3} &= S=\frac{b_0^9 r\left(1-r\right)\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)}{\left(b_0^3 r\left(1-r\right)\right)^3} \\ 0 &= r^6-246r^5+(3487-S)\,r^4+(2S-10452)\,r^3+(3487-S)\,r^2-246r+1 \end{aligned}$$

Subtracting the polynomial with Q from the polynomial with S gives

$$0=216r^4+(S-Q-3184)\,r^3+(9392+2Q-2S)\,r^2+(S-Q-3184)\,r+216.$$

Multiply this by $(1-R)$, multiply the polynomial based on R by 216, subtract one from the other and divide by r (which is nonzero) to obtain an expression that simplifies to

$$0=Nr^2-\left(2\left(1-R\right)\left(6320+S-Q\right)+31104\right)r+N$$

where $N=(1-R)\,(1136+S-Q)+7776$, which after substituting in for R, S, and Q becomes

$$N=\frac{15552r\left(r+1\right)^4}{\left(r^2-10r+1\right)^2\left(1-r\right)^2}$$

The denominator of this expression for N is not equal to zero, because that would imply $s=0$, and we are currently specifically considering the case where $s\neq 0$ (having already analyzed the case where $s = 0$). Also $N \neq 0$ because $r \neq 0$, and $r \neq -1$. We therefore have $0 = r^2 - \gamma r + 1$ where $\gamma = \left(2\left(1-R\right)\left(6320 + S - Q\right) + 31104\right)/N$, which is identified because all of its components are identified.

We have now shown that $0 = r^2 - \gamma r + 1$ where γ is identified. This equation says that $\gamma = r + r^{-1} = \left[p_0/\left(1-p_0\right)\right] + \left[\left(1-p_0\right)/p_0\right]$. Whatever value p_0 takes on between zero and one makes this expression for γ greater than or equal to two. The equation $0 = r^2 - \gamma r + 1$ has solutions

$$r = \frac{1}{2}\gamma + \frac{1}{2}\sqrt{\gamma^2 - 4} \quad \text{and} \quad r = \frac{1}{\frac{1}{2}\gamma + \frac{1}{2}\sqrt{\gamma^2 - 4}}$$

with $\gamma^2 \geq 4$, so one of these solutions must be the true value of r. Given r, we can then solve for b_0 by $b_0 = e_3^{1/3}\left(r\left(1-r\right)\right)^{1/3}$. Recall that $r = p_0/p_1$. If we exchanged b_0 with b_1 and exchanged p_0 with p_1 everywhere, all of the above equations would still hold. It follows that one of the above two values of r must equal p_0/p_1, and the other equals p_1/p_0. The former when substituted into $e_3\left(r\left(1-r\right)\right)$ will yield b_0^3 and the latter must yield b_1^3. Without loss of generality imposing the constraint $b_0 < 0 < b_1$ shows that the correct solution for r will be the one that satisfies $e_3\left(r\left(1-r\right)\right) < 0$, and so r and b_0 is identified. The remainder of the distribution of V is then given by $p_0 = r/\left(1+r\right)$, $p_1 = 1/\left(1+r\right)$, and $b_1 = -b_0 r$.

Finally, we show identification of the distribution of U. For any random variable Z, let F_Z denote the marginal cumulative distribution function of Z. By the probability mass function of the V distribution, $F_\varepsilon\left(\varepsilon\right) = \left(1-p\right)F_U\left(\varepsilon - b_1\right) + pF_U\left(\varepsilon - b_0\right)$. Letting $\varepsilon = u + \left(b_0 - b_1\right)k - b_0$ and rearranging gives

$$F_U\left(u + \left(b_0 - b_1\right)k\right) = \frac{1}{p}F_\varepsilon\left(u + b_0 + \left(b_0 - b_1\right)k\right) - \frac{1-p}{p}F_U\left(u + \left(b_0 - b_1\right)\left(k+1\right)\right)$$

so for positive integers K, $F_U\left(u\right) = R_K + \sum_{k=0}^{K-1}\left(\frac{1-p}{p}\right)^k \frac{1}{p}F_\varepsilon\left(u + b_0 + \left(b_0 - b_1\right)k\right)$ where the remainder term $R_K = r^{-K}F_U\left(u + \left(b_0 - b_1\right)K\right) \leq r^{-K}$. If $r > 1$ then $R_k \to 0$ as $K \to \infty$, so $F_U\left(u\right)$ is identified by

$$F_U\left(u\right) = \sum_{k=0}^{\infty}\left(\frac{1-p}{p}\right)^k \frac{1}{p}F_\varepsilon\left(u + b_0 + \left(b_0 - b_1\right)k\right) \tag{23}$$

since all the terms on the right of this expression are identified, given that the distributions of ε and of V are identified. If $r < 1$, then exchange the roles of b_0 and b_1 (e.g., start by letting $\varepsilon = u + (b_1 - b_0)\, k - b_1)$ which will correspondingly exchange p and $1-p$ to obtain $F_U(u) = \sum_{k=0}^{\infty} \left(\frac{p}{1-p}\right)^k \frac{1}{1-p} F_\varepsilon (u + b_1 + (b_1 - b_0)\, k)$, where now the remainder term was $R_K = r^K F_U (u + (b_1 - b_0)\, K) \leq r^K \to 0$ as $K \to \infty$ since now $r < 1$. The case of $r = 0$ is ruled out, since that is equivalent to $p = 1/2$.

PROOF of Proposition 1: $Y = h + V + U$ and independence of U and V implies that

$$E[\exp(\tau (Y - h))] = E[\exp(\tau V)]\, E[\exp(\tau U)]$$

Now $E[\exp(\tau V)] = p \exp(\tau b_0) + (1-p)\exp(\tau b_1)$. Define $\alpha_\tau = (1-p)\, E\left(e^{\tau U}\right)$. By symmetry of U, $\alpha_\tau = \alpha_{-\tau}$. These equations with $r = p/(1-p)$ and $b_1 = b_0 p/(p-1)$ give equations (20) and (21).

PROOF of Theorem 2: By the probability mass function of the V distribution, $F_\varepsilon(\varepsilon) = (1-p)\, F_U(\varepsilon - b_1) + p F_U(\varepsilon - b_0)$. Evaluating this expression at $\varepsilon = u + b_1$ gives

$$F_\varepsilon(u + b_1) = (1-p)\, F_U(u) + p F_U(u + b_1 - b_0) \tag{24}$$

and evaluating at $\varepsilon = -u + b_0$ gives $F_\varepsilon(-u + b_0) = (1-p)\, F_U(-u - b_1 + b_0) + p F_U(-u)$. Apply symmetry of U which implies $F_U(u) = 1 - F_U(-u)$ to this last equation to obtain

$$F_\varepsilon(-u + b_0) = (1-p)\,[1 - F_U(U + b_1 - b_0)] + p\,[1 - F_U(u)] \tag{25}$$

Equations (24) and (25) are two equations in the two unknowns $F_U(U + b_1 - b_0)$ and $F_U(U)$. Solving for $F_U(U)$ gives $F_U(U) = \Psi(U)$ with $\Psi(U)$ given by equation (11). It follows from symmetry of U that $F_U(U)$ must also equal $1 - \Psi(-U)$, which gives equation (12).

PROOF of Corollary 1: First identify $h(x)$ by $h(x) = E(Y \mid X = x)$, since $E(Y - h(X) \mid X = x) =$
$E(V + U \mid X = x)$
$= E(V + U) = 0$. Next define $\varepsilon = Y - h(X)$ and then the rest of the proof is identical to the proof of Theorem 1.

PROOF of Corollary 2: Define $h(x) = E(Y \mid X)$ and $\varepsilon = Y - h(X)$. Then $h(x)$ and the distribution of ε conditional upon X is identified and $E(\varepsilon \mid X) = 0$. Define $V = g(X, D^*) - h(X)$ and let $b_d(X) = g(X, d) - h(X)$ for $d = 0, 1$. Then $\varepsilon = V + U$, where V (given X) has the distribution with support equal to the two values $b_0(X)$ and

$b_1(X)$ with probabilities $p(X)$ and $1-p(X)$, respectively. Also U and ε have mean zero given X so $E\left(V \mid X\right) = 0$. Applying Theorem 1 separately for each value x on the support of X shows that $b_0(x)$, $b_1(x)$, $p\left(x\right)$, and the conditional distribution of U given $X = x$ are identified for each such x, and it follows that the function $g(x, d)$ is identified by $g(x, d) = b_d(x) + h\left(x\right)$.

Table 1: p = .6

PARAMETER	b1	b0	p	u2	u4	u6
TRUE	2.500	-1.667	0.600	1.000	3.000	15.000
MEDIAN	2.491	-1.657	0.601	0.983	2.786	12.309
MEAN	2.374	-1.578	0.586	1.234	4.846	34.142
STDDEV	0.506	0.347	0.099	0.925	7.946	90.287
ROOT MSE	0.522	0.358	0.100	0.954	8.154	92.250
MEDIAN ABS ERR	0.068	0.057	0.013	0.049	0.380	3.927
MEAN ABS ERR	0.201	0.146	0.034	0.302	2.399	25.013
25% QUANTILE	2.422	-1.712	0.588	0.942	2.518	9.973
75% QUANTILE	2.561	-1.599	0.614	1.034	3.106	15.210

Table 2: p = .8

PARAMETER	b1	b0	p	u2	u4	u6
TRUE	5.000	-1.250	0.800	1.000	3.000	15.000
MEDIAN	4.955	-1.242	0.799	1.007	2.858	13.222
MEAN	4.601	-1.267	0.756	1.018	3.228	17.151
STDDEV	1.235	0.402	0.153	0.368	2.657	26.392
ROOT MSE	1.298	0.403	0.159	0.368	2.665	26.466
MEDIAN ABS ERR	0.094	0.065	0.009	0.065	0.337	3.106
MEAN ABS ERR	0.459	0.175	0.054	0.164	0.884	7.905
25% QUANTILE	4.871	-1.305	0.789	0.946	2.548	10.549
75% QUANTILE	5.036	-1.178	0.808	1.078	3.169	15.951

Table 3: p = .95

PARAMETER	b1	b0	p	u2	u4	u6
TRUE	20.000	-1.053	0.950	1.000	3.000	15.000
MEDIAN	19.948	-1.051	0.950	0.988	2.890	14.997
MEAN	19.852	-1.051	0.947	0.987	2.852	14.501
STDDEV	1.404	0.161	0.043	0.122	0.529	16.208
ROOT MSE	1.411	0.161	0.043	0.123	0.549	16.208
MEDIAN ABS ERR	0.142	0.103	0.005	0.083	0.315	0.010
MEAN ABS ERR	0.265	0.123	0.008	0.096	0.397	4.348
25% QUANTILE	19.816	-1.151	0.945	0.903	2.527	14.946
75% QUANTILE	20.083	-0.943	0.955	1.064	3.121	15.002

Nonparametric Identification of a Binary Random Factor in Cross Section Data - Supplemental Appendix

Yingying Dong and Arthur Lewbel*

California State University Fullerton and Boston College

July 2010

Abstract

This supplemental appendix to,"Nonparametric Identification of a Binary Random Factor in Cross Section Data" provides:

1. An empirical application of the estimator

2. A more detailed proof of the paper's main Theorem, filling in many tedious algebra steps.

3. Limiting distribution theory for the local GMM estimator that is summarized in the paper.

Identifying moments:

$$v_d = E\left(V^d\right) = b_0^d p + \left(\frac{b_0 p}{p-1}\right)^d (1-p)\,. \tag{1}$$

$$E\left(Y-h\right) = 0 \tag{2}$$

$$E\left(\left(Y-h\right)^2 - \left(v_2+u_2\right)\right) = 0 \tag{3}$$

$$E\left(\left(Y-h\right)^3 - v_3\right) = 0 \tag{4}$$

$$E\left(\left(Y-h\right)^4 - \left(v_4+6v_2u_2+u_4\right)\right) = 0 \tag{5}$$

[*] Corresponding Author: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu,http://www2.bc.edu/~lewbel/

$$E\left((Y-h)^5-(v_5+10v_3u_2)\right)=0 \tag{6}$$

$$E\left((Y-h)^6-(v_6+15v_4u_2+15v_2u_4+u_6)\right)=0 \tag{7}$$

$$E\left((Y-h)^7-(v_7+21v_5u_2+35v_3u_4)\right)=0 \tag{8}$$

$$E\left((Y-h)^9-(v_9+36v_7u_2+126v_5u_4+84v_3u_6)\right)=0 \tag{9}$$

Density of U estimation:

$$\widehat{\psi}\left(u\right)=\frac{-\widehat{f}_{\varepsilon}\left(-u+\widehat{b}_0\right)\widehat{p}+\widehat{f}_{\varepsilon}\left(u+\widehat{b}_1\right)(1-\widehat{p})}{1-2\widehat{p}}, \tag{10}$$

$$\widehat{f_U}\left(u\right)=\frac{\widehat{\psi}\left(u\right)+\widehat{\psi}\left(-u\right)}{2}, \tag{11}$$

where $\widehat{f_{\varepsilon}}\left(\varepsilon\right)$ is a kernel density or other estimator of $f_{\varepsilon}\left(\varepsilon\right)$, constructed using data $\widehat{\varepsilon}_i=Y_i-\widehat{h}$ for $i=1,...n$.

1 A Parametric U Comparison

It might be useful to construct parametric estimates of the model, which could for example provide reasonable starting values for the GMM estimation. The parametric model we propose for comparison assumes that U is normal with mean zero and standard deviation s.

When U is normal the distribution of Y is finitely parameterized, and so can be estimated directly by maximum likelihood. The log likelihood function is given by

$$\sum_{i=1}^{n}\ln\left(\frac{p}{s\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{Y_i-h-b_0}{s}\right)^2\right]+\frac{1-p}{s\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{Y_i-h-\frac{b_0p}{p-1}}{s}\right)^2\right]\right). \tag{12}$$

Maximizing this log likelihood function provides estimates of h, b_0, p, and s. As before, an estimate of b_1 would be given by $\widehat{b}_1 = \widehat{b}_0\widehat{p}/(\widehat{p}-1)$. Further, if U is normal then $u_2 = s^2$, $u_4 = 3s^2$, and $u_6 = 15s^2$. These estimates can be compared to the GMM estimates, which should be the same if the true distribution of U is indeed normal.

2 An Empirical Application: World Income Distribution

A large literature exists regarding the distribution of income across countries, much of which deals with the question of convergence, that is, whether poorer countries are catching up with richer countries as a result of increases in globalization of trade and diffusion of technology.

To measure the extent of convergence, if any, we propose a simple descriptive model of the income distribution across countries. Assume that there exist two types of countries, i.e., poor versus rich, or less developed versus more developed countries. Let I_{ti} denote the per capita income or GDP of country i in time t, and define Y_{ti} to be either income levels $Y_{ti} = I_{ti}$, or income shares $Y_{ti} = I_{ti}/\left(\sum_{i=1}^{n} I_{ti}\right)$. Assume that a poor country's income in year t is given by $Y_{ti} = g_{t0} + U_{ti}$, while that of a wealthy country is given by $Y_{ti} = g_{t1} + U_{ti}$, where g_{t0} and g_{t1} are the mean income levels or mean shares for poor and rich countries, respectively, and U_{ti} is an individual country's deviation from its group mean. Here U_{ti} embodies both the relative ranking of country i within its (poor or rich) group, and may also include possible measurement errors in Y_{ti}. We assume that the distribution of U_{ti} is symmetric and mean zero with a probability density function f_{tu}.

Let $h_t = E_t(Y)$ be the mean income or income share for the whole population of countries in year t. Then the income measure for country i in year t can be written as $Y_{ti} = h_t + V_{ti} + U_{ti}$, where V_{ti} is the deviation of rich or poor countries' group mean from the grand mean h_t. Then V_{ti} equals $b_{t0} = g_{t0} - h_t$ with probability p_t and V_{ti} equals $b_{t1} = g_{t1} - h_t$ with probability $1 - p_t$, so p_t is the fraction of countries that are in the poor group in year t, and $b_{t1} - b_{t0}$ is the difference in mean income or income shares between poor and wealthy countries.

Objections can be easily raised to this simplistic model, e.g., that other indicators in addition to income exist for grouping countries, that countries could be divided into more than two groups, and that there is not a strong economic argument for why the distribution of incomes around group means should be symmetric and the same for both groups. One could respond that it is common to dichotomize the world into groups of rich (well developed) and poor (less developed) countries, that Gibrat's law within groups could generate the required symmetry, and that the shape of the world income distribution suggests at least rough appropriateness of the model (including possible bimodality of Y with estimates of the U distribution close to normal). Still, given these valid concerns, we interpret our model as primarily descriptive rather than structural. Our main goal is to verify that the polynomial moments we use for identification and estimation can produce reasonable estimates with real data and small sample sizes.

Though simple, our model provides measures of a few different possible types of convergence. Having p_t decrease over time would indicate that on average countries are leaving the poor group and joining the set of wealthy nations. A finding that $b_{t1} - b_{t0}$ decreases over time would mean that the differences between rich and poor nations are diminishing, and a finding that the spread (e.g. the variance) of the density f_{tu} decreases over time would mean that there is convergence within but not necessarily across the poor and rich groups.

A feature of this model is that it does not require arbitrarily choosing a threshold level of Y to demarcate the line between rich and poor countries, and so avoids this potential source of misspecification. This model also allows for the possibility that a poor country has higher income than some wealthy country in a given time period due to random factors (e.g., natural disaster in a wealthy country i, implying a low draw of U_{ti} in time t). More generally, the model does not require specifying or estimating the group to which each country belongs.

Bianchi (1997) applies bimodality tests to the distribution of income across countries over time, to address questions regarding evidence for convergence. Bimodality versus unimodality of Y might be interpreted as evidence in favor of a 'two group' model, though note that even if U is unimodal, e.g., normal, then Y can be either unimodal or bimodal (with possibly large differences in the heights of the two modes), depending on p and on the magnitudes of b_0 and b_1. The density for Y can also be quite skewed, even though U is symmetric.

For comparison we apply our model using the same data as Bianchi, which consists of I_{it} defined as annual per capita GDP in constant U.S. dollars for 119 countries, measured in 1970, 1980 and 1989.

Table 1: Estimates based on the GDP per capita level data (in 10,000 1985 dollars)

		p	b0	b1	b1-b0	h	u2	u4	u6
	GMM	.8575	-.1105	.6648	.7753	.3214	.0221	.0001$^{\$}$	.0024
1970		(.0352)	(.0244)	(.0664)	(.0590)	(.0284)	(.0042)	(.0002)	(.0009)
	MLE	.8098	-.1334	.5679	.7013	.3213	.0199		
		(.0362)	(.0260)	(.0487)	(.0477)	(.0280)	(.0031)		
	GMM	.8081	-.1722	.7252	.8974	.4223	.0294	.0016	.0017*
1980		(.0371)	(.0322)	(.0579)	(.0491)	(.0351)	(.0043)	(.0004)	(.0007)
	MLE	.8070	-.1692	.7077	.8769	.4222	.0350		
		(.0393)	(.0345)	(.0600)	(.0544)	(.0372)	(.0048)		
	GMM	.8125	-.2114	.9159	1.1273	.4804	.0384	.0051	.0028$^{\$}$
1989		(.0380)	(.0424)	(.1022)	(.1111)	(.0439)	(.0118)	(.0104)	(.0448)
	MLE	.7948	-.2192	.8491	1.0683	.4805	.0489		
		(.0393)	(.0413)	(.0754)	(.0679)	(.0441)	(.0076)		

Note: $^{\$}$ not significant; * significant at the 5% level; all the others are significant at the 1% level. Standard errors are in parentheses.

Table 2: Estimates based on the scaled GDP per capita share data

		p	b0	b1	b1-b0	h	u2	u4	u6
1970	GMM	.8619	-.1392	.8682	1.0074	.4206	.0417	.0039$	$.0057^{\$}$
		(.0361)	(.0332)	(.1009)	(.0985)	(.0380)	(.0089)	(.0068)	(.0063)
	MLE	.8098	-.1744	.7425	.9169	.4202	.0340		
		(.0383)	(.0352)	(.0670)	(.0629)	(.0377)	(.0053)		
1980	GMM	.8080	-.1715	.7217	.8932	.4202	.0291	.0016	.0017
		(.0374)	(.0334)	(.0560)	(.0497)	(.0364)	(.0041)	(.0004)	(.0006)
	MLE	.8070	-.1684	.7043	.8727	.4202	.0347		
		(.0373)	(.0322)	(.0570)	(.0508)	(.0353)	(.0045)		
1989	GMM	.8117	-.1848	.7964	.9812	.4203	.0316	.0023	$.0020^{*}$
		(.0360)	(.0344)	(.0609)	(.0518)	(.0388)	(.0049)	(.0007)	(.0009)
	MLE	.7948	-.1916	.7424	.934	.4202	.0374		
		(.0387)	(.0355)	(.0655)	(.0589)	(.0395)	(.0058)		

Note: $^{\$}$ not significant; *significant at the 5% level; all the others are significant at the 1% level. Standard errors are in parentheses.

For each of the three years of data we provide two different estimates, labeled GMM and MLE in Tables 1 and 2. GMM is based on the identifying polynomial moments (2) to (9) (after substituting in equation (1)), while MLE is a maximum likelihood estimator that maximizes (12) assuming that U is normal.

Table 1 reports results based on per capita levels, $Y_{ti} = I_{ti}/10,000$, while Table 2 is based on scaled shares, $Y_{ti} = 50I_{ti}/\left(\sum_{i=1}^{n} I_{ti}\right)$. We scale by 10,000 in Table 1 and by 50 in Table 2 to put the Y_{ti} data in a range between zero and two in each case. These scalings are theoretically irrelevant, but in practice help ensure that the matrices involved in estimation (particularly the high order polynomial terms in the estimated second stage GMM weighting matrix) are numerically well conditioned despite computer round off error.

In both Tables 1 and 2, and in all three years, the GMM and maximum likelihood estimates are roughly comparable, for the most part lying within about 10% of each other. Looking across years, both Tables tell similar stories in terms of percentages of poor countries. Using either levels or shares, by GMM p is close to .86 in 1970, and close to .81 in 1980 and 1989, showing a decline in the number of poor countries in the 1970's, but no further decline in the 1980's. In contrast, MLE shows p close to .81 in all years. The average difference between rich and poor, $b_1 - b_0$, increases steadily over time in the levels data, but this may be due in part to the growth of average income over time, given by h. Share data scales out this income growth over time. Estimates based on shares in Table 2 show that $b_1 - b_0$ decreased by a small amount in the 1970's, but then increased again in the 1980's, so by this measure there is no clear evidence of convergence or divergence.

Figure 1 shows $\widehat{f}_u$, the estimated density of U, given by equation (11) using the GMM estimates from Table 2 in 1970. Graphs of other years are very similar, so to save space we do not include them here. This estimated density is compared to a normal density with the same mode, $\widehat{f}_u(0)$. It follows that this normal density has standard deviation $(2\pi)^{-1/2}\left[\widehat{f}_u(0)\right]^{-1}$. With the same central tendency given by construction, these two densities can be compared for differences in dispersion and tail behaviors. As Figure 1 shows, the semiparametric $\widehat{f}_u$ matches the normal density rather closely except near the tails of its distribution where data are sparse. Also shown in Figure 1 is the maximum likelihood estimate of f_u, which assumes U is normal. Although close to normal in shape, the semiparametric $\widehat{f}_u$ appears to have a larger variance than the maximum likelihood estimate. The graphs of $\widehat{f}_u$ in other years are very similar, and they along with the variance estimates in Table 2 show no systematic trends in the dispersion of U over time, and hence no evidence of income convergence within groups of rich or poor countries.

In this analysis of U, note that Y is by construction nonnegative so U cannot literally be normal; however, the value of U where $Y = h + V + U$ crosses zero is far out in the left tail of the U distribution (beyond the values graphed in Figure 1), so imposing the constraint on U that Y be nonnegative (e.g., making the parametric comparison U a truncated normal) would have no discernable impact on the resulting estimates.

In addition to levels I_{ti} and shares $I_{ti}/\left(\sum_{i=1}^{n} I_{ti}\right)$, Bianchi (1997) also considers logged data, but finds that the log transformation changes the shape of the Y_{ti} distribution in a way that obscures bimodality. We found similar results, in that with logged data our model yields estimates of p close to .5, which is basically ruled out by our model, as $p = .5$ would make V be symmetric and hence unidentifiable relative to U. As noted earlier, one can readily tell a priori if p is close to .5, because this can happen only if the observed Y distribution is itself close to symmetric.

3 Detailed Proof of Theorem 1

PROOF of Theorem 1: First identify h by $h = E(Y)$, since V and U are mean zero. Then the distribution of ε defined by $\varepsilon = Y - h$ is identified, and $\varepsilon = U + V$. Define $e_d = E\left(\varepsilon^d\right)$ and $v_d = E\left(V^d\right)$.

Now evaluate e_d for integers $d \leq 9$. These e_d exist by assumption, and are identified because the distribution of ε is identified. The first goal will be to obtain expressions for v_d in terms of e_d for various values of d. Using independence of V and U, the fact that both are mean zero, and U being symmetric we have

$$\begin{aligned}
E\left(\varepsilon^2\right) &= E\left(V^2 + 2VU + U^2\right) \\
e_2 &= v_2 + E\left(U^2\right) \\
E\left(U^2\right) &= e_2 - v_2
\end{aligned}$$

$$E\left(\varepsilon^3\right) = E\left(V^3 + 3V^2U + 3VU^2 + U^3\right)$$

$$e_3 = v_3$$

$$\begin{aligned} E\left(\varepsilon^4\right) &= E\left(V^4 + 4V^3U + 6V^2U^2 + 4VU^3 + U^4\right) \\ e_4 &= v_4 + 6v_2E\left(U^2\right) + E\left(U^4\right) \\ E\left(U^4\right) &= e_4 - v_4 - 6v_2E\left(U^2\right) \\ &= e_4 - v_4 - 6v_2\left(e_2 - v_2\right) \\ E\left(U^4\right) &= e_4 - v_4 - 6v_2e_2 + 6v_2^2 \end{aligned}$$

$$\begin{aligned} E\left(\varepsilon^5\right) &= E\left(V^5 + 5V^4U + 10V^3U^2 + 10V^2U^3 + 5VU^4 + U^5\right) \\ e_5 &= v_5 + 10v_3E\left(U^2\right) = v_5 + 10v_3\left(e_2 - v_2\right) \\ e_5 &= v_5 + 10e_3e_2 - 10e_3v_2 \\ e_5 - 10e_3e_2 &= v_5 - 10e_3v_2 \end{aligned}$$

Define $s = e_5 - 10e_3e_2$, and note that s depends only on identified objects and so is identified. Then $s = v_5 - 10e_3v_2$,

$$\begin{aligned} E\left(\varepsilon^6\right) &= E\left(V^6 + 6V^5U + 15V^4U^2 + 20V^3U^3 + 15V^2U^4 + 6VU^5 + U^6\right) \\ e_6 &= v_6 + 15v_4E\left(U^2\right) + 15v_2E\left(U^4\right) + E\left(U^6\right) \\ E\left(U^6\right) &= e_6 - v_6 - 15v_4E\left(U^2\right) - 15v_2E\left(U^4\right) \\ &= e_6 - v_6 - 15v_4\left(e_2 - v_2\right) - 15v_2\left(e_4 - v_4 - 6v_2e_2 + 6v_2^2\right) \\ &= e_6 - v_6 - 15e_2v_4 - 15e_4v_2 + 30v_2v_4 - 90v_2^3 + 90e_2v_2^2 \end{aligned}$$

$$E\left(\varepsilon^7\right) = E\left(V^7 + 7V^6U + 21V^5U^2 + 35V^4U^3 + 35V^3U^4 + 21V^2U^5 + 7VU^6 + U^7\right)$$

$$\begin{aligned} e_7 &= v_7 + 21 v_5 E\left(U^2\right) + 35 v_3 E\left(U^4\right) \\ e_7 &= v_7 + 21 v_5 \left(e_2 - v_2\right) + 35 v_3 \left(e_4 - v_4 - 6 v_2 e_2 + 6 v_2^2\right) \end{aligned}$$

plug in $v_5 = s + 10 e_3 v_2$ and $v_3 = e_3$ and expand:

$$\begin{aligned} e_7 &= v_7 + 21\left(s + 10 e_3 v_2\right)\left(e_2 - v_2\right) + 35 e_3 \left(e_4 - v_4 - 6 v_2 e_2 + 6 v_2^2\right) \\ &= v_7 + 21 s e_2 - 21 s v_2 + 35 e_3 e_4 - 35 e_3 v_4 \end{aligned}$$

Bring terms involving identified objects e_d and s left:

$$e_7 - 21 s e_2 - 35 e_3 e_4 = v_7 - 35 e_3 v_4 - 21 s v_2.$$

Define $q = e_7 - 21 s e_2 - 35 e_3 e_4$ and note that q depends only on identified objects and so is identified. Then

$$q = v_7 - 35 e_3 v_4 - 21 s v_2.$$

Next consider e_9.

$$\begin{aligned} E\left(\varepsilon^9\right) &= E\left(\begin{array}{c} V^9 + 9V^8U + 36V^7U^2 + 84V^6U^3 + 126V^5U^4 + \\ 126V^4U^5 + 84V^3U^6 + 36V^2U^7 + 9VU^8 + U^9 \end{array}\right) \\ e_9 &= v_9 + 36 v_7 E\left(U^2\right) + 126 v_5 E\left(U^4\right) + 84 v_3 E\left(U^6\right) \\ e_9 &= v_9 + 36 v_7 \left(e_2 - v_2\right) + 126 v_5 \left(e_4 - v_4 - 6 v_2 e_2 + 6 v_2^2\right) \\ &\quad + 84 v_3 \left(e_6 - v_6 - 15 e_2 v_4 - 15 e_4 v_2 + 30 v_2 v_4 - 90 v_2^3 + 90 e_2 v_2^2\right) \end{aligned}$$

Use q and s to substitute out $v_7 = q + 35 e_3 v_4 + 21 s v_2$ and $v_5 = s + 10 e_3 v_2$, and use $v_3 = e_3$ to get

$$e_9 = v_9 + 36\left(q + 35 e_3 v_4 + 21 s v_2\right)\left(e_2 - v_2\right) + 126\left(s + 10 e_3 v_2\right)\left(e_4 - v_4 - 6 v_2 e_2 + 6 v_2^2\right)$$

$$+84e_3\left(e_6 - v_6 - 15e_2v_4 - 15e_4v_2 + 30v_2v_4 - 90v_2^3 + 90e_2v_2^2\right)$$

Expand and bring terms involving identified objects e_d, s, and q to the left:

$$e_9 - 36qe_2 - 126se_4 - 84e_3e_6 = v_9 - 36qv_2 - 126sv_4 - 84e_3v_6$$

Define $w = e_9 - 36qe_2 - 126se_4 - 84e_3e_6$ and note that w depends only on identified objects and so is identified. Then

$$w = v_9 - 36qv_2 - 126sv_4 - 84e_3v_6$$

Summarizing, we have w, s, q, e_3 are all identified and

$$\begin{aligned} e_3 &= v_3 \\ s &= v_5 - 10e_3v_2 \\ q &= v_7 - 35e_3v_4 - 21sv_2 \\ w &= v_9 - 84e_3v_6 - 126sv_4 - 36qv_2. \end{aligned}$$

Now V only takes on two values, so let V equal b_0 with probability p_0 and b_1 with probability p_1. Probabilities sum to one, so $p_1 = 1 - p_0$. Also, $E\left(V\right) = b_0p_0 + b_1p_1 = 0$ because $\varepsilon = V + U$ and both ε and U have mean zero, so $b_1 = -b_0p_0/\left(1 - p_0\right)$. Let $r = p_0/p_1 = p_0/\left(1 - p_0\right)$, so

$$p_0 = r/\left(1 + r\right), \qquad p_1 = 1/\left(1 + r\right), \qquad b_1 = -b_0r,$$

and for any integer d

$$v_d = b_0^d p_0 + b_1^d p_1 = b_0^d \left(p_0 + (-r)^d p_1\right) = b_0^d \frac{r + (-r)^d}{1 + r}$$

so in particular

$$\begin{aligned}
v_2 &= b_0^2 r \\
v_3 &= b_0^3 r (1 - r) \\
v_4 &= b_0^4 r \left(r^2 - r + 1\right) \\
v_5 &= b_0^5 r (1 - r) \left(r^2 + 1\right) \\
v_6 &= b_0^6 \frac{r + (-r)^6}{1 + r} = b_0^6 r \left(r^4 - r^3 + r^2 - r + 1\right) \\
v_7 &= b_0^7 r (1 - r) \left(r^4 + r^2 + 1\right) \\
v_9 &= b_0^9 \frac{r + (-r)^9}{1 + r} = b_0^9 r (1 - r) \left(r^2 + 1\right) \left(r^4 + 1\right)
\end{aligned}$$

Substituting these v_d expressions into the expression for e_3, s, q, and w gives $e_3 = b_0^3 r (1 - r)$,

$$\begin{aligned}
s &= b_0^5 r (1 - r) \left(r^2 + 1\right) - 10 b_0^3 r (1 - r) b_0^2 r \\
s &= b_0^5 r (1 - r) \left(r^2 - 10r + 1\right)
\end{aligned}$$

$$\begin{aligned}
q &= v_7 - 35 e_3 v_4 - 21 s v_2 \\
&= b_0^7 r (1 - r) \left(r^4 + r^2 + 1\right) - 35 b_0^3 r (1 - r) b_0^4 r \left(r^2 - r + 1\right) - 21 b_0^5 r (1 - r) \left(r^2 - 10r + 1\right) b_0^2 r \\
q &= b_0^7 r (1 - r) \left(r^4 - 56r^3 + 246r^2 - 56r + 1\right)
\end{aligned}$$

$$\begin{aligned}
w &= v_9 - 84e_3v_6 - 126sv_4 - 36qv_2 \\
&= \left(\begin{array}{c}
b_0^9 r\,(1-r)\,(r^2+1)\,(r^4+1) - 84\,(b_0^3 r\,(1-r))\,(b_0^6 r\,(r^4-r^3+r^2-r+1)) \\
-126\,(b_0^5 r\,(1-r)\,(r^2-10r+1))\,(b_0^4 r\,(r^2-r+1)) \\
-36\,(b_0^7 r\,(1-r)\,(r^4-56r^3+246r^2-56r+1))\,b_0^2 r
\end{array}\right) \\
w &= b_0^9 r\,(1-r)\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)
\end{aligned}$$

Summarizing the results so far we have

$$\begin{aligned}
e_3 &= b_0^3 r\,(1-r) \\
s &= b_0^5 r\,(1-r)\left(r^2-10r+1\right) \\
q &= b_0^7 r\,(1-r)\left(r^4-56r^3+246r^2-56r+1\right) \\
w &= b_0^9 r\,(1-r)\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)
\end{aligned}$$

These are four equations in the two unknowns b_0 and r. We require all four equations for identification, and not just two or three of them, because these are polynomials in r and so have multiple roots. We will now show that these four equations imply that $r^2 - \gamma r + 1 = 0$, where γ is finite and identified.

First we have $e_3 = v_3 \neq 0$ and $r \neq 1$ by asymmetry of V. Also $r \neq 0$ because then V would only have one point of support instead of two, and these together imply by $e_3 = b_0^3 r\,(1-r)$ that $b_0 \neq 0$. Applying these results to the s equation shows that if s (which is identified) is zero then $r^2 - 10r + 1 = 0$, and so in that case γ is identified. So now consider the case where $s \neq 0$.

Define $R = qe_3/s^2$, which is identified because its components are identified. Then

$$\begin{aligned}
R &= \frac{b_0^7 r\,(1-r)\,(r^4-56r^3+246r^2-56r+1)\,b_0^3 r\,(1-r)}{b_0^5 r\,(1-r)\,(r^2-10r+1)\,b_0^5 r\,(1-r)\,(r^2-10r+1)} \\
&= \frac{r^4-56r^3+246r^2-56r+1}{\left(r^2-10r+1\right)^2}
\end{aligned}$$

So

$$\begin{aligned}
0 &= \left(r^4 - 56r^3 + 246r^2 - 56r + 1\right) - \left(r^2 - 10r + 1\right)^2 R \\
0 &= (1 - R)\, r^4 + (-56 + 20R)\, r^3 + (246 - 102R)\, r^2 + (-56 + 20R)\, r + (1 - R)
\end{aligned}$$

Which yields a fourth degree polynomial in r. If $R = 1$, then (using $r \neq 0$) this polynomial reduces to the quadratic $0 = r^2 - 4r + 1$, so in this case $\gamma = -4$ is identified. Now consider the case where $R \neq 1$.

Define $Q = s^3/e_3^5$ which is identified because its components are identified. Then

$$\begin{aligned}
Q &= \frac{\left(b_0^5 r\,(1-r)\,(r^2 - 10r + 1)\right)^3}{\left(b_0^3 r\,(1-r)\right)^5} = \frac{\left(r^2 - 10r + 1\right)^3}{\left(r\,(1-r)\right)^2} \\
0 &= \left(r^2 - 10r + 1\right)^3 - \left(r\,(1-r)\right)^2 Q \\
0 &= r^6 - 30r^5 + (303 - Q)\, r^4 + (2Q - 1060)\, r^3 + (303 - Q)\, r^2 - 30r + 1
\end{aligned}$$

which is a sixth degree polynomial in r. Also define $S = w/e_3^2$ which is identified because its components are identified. Then

$$\begin{aligned}
\frac{w}{e_3^3} &= S = \frac{b_0^9 r\,(1-r)\,(r^6 - 246r^5 + 3487r^4 - 10452r^3 + 3487r^2 - 246r + 1)}{\left(b_0^3 r\,(1-r)\right)^3} \\
S &= \frac{(r^6 - 246r^5 + 3487r^4 - 10452r^3 + 3487r^2 - 246r + 1)}{\left(r\,(1-r)\right)^2} \\
0 &= \left(r^6 - 246r^5 + 3487r^4 - 10452r^3 + 3487r^2 - 246r + 1\right) - \left(r\,(1-r)\right)^2 S \\
0 &= r^6 - 246r^5 + (3487 - S)\, r^4 + (2S - 10452)\, r^3 + (3487 - S)\, r^2 - 246r + 1
\end{aligned}$$

which is another sixth degree polynomial in r. Subtracting the second of these sixth degree polynomials from the other and dividing the result by r gives the fourth order polynomial:

$$0 = 216r^4 + (S - Q - 3184)\, r^3 + (9392 + 2Q - 2S)\, r^2 + (S - Q - 3184)\, r + 216.$$

Multiply this fourth order polynomial by $(1 - R)$, multiply the previous fourth order polynomial by 216, subtract one from the other. and divide by r to obtain a quadratic in r:

$$\begin{aligned} 0 \;=\;& 216\,(1-R)\, r^4 + (1-R)\,(S-Q-3184)\, r^3 + (1-R)\,(9392+2Q-2S)\, r^2 \\ & + (1-R)\,(S-Q-3184)\, r + 216\,(1-R) - 216\,(1-R)\, r^4 - 216\,(-56+20R)\, r^3 \\ & -216\,(246-102R)\, r^2 - 216\,(-56+20R)\, r - 216\,(1-R) \end{aligned}$$

$$\begin{aligned} 0 \;=\;& ((1-R)\,(S-Q-3184) - 216\,(-56+20R))\, r^3 \\ & + ((1-R)\,(9392+2Q-2S) - 216\,(246-102R))\, r^2 \\ & + ((1-R)\,(S-Q-3184) - 216\,(-56+20R))\, r \end{aligned}$$

$$\begin{aligned} 0 \;=\;& ((1-R)\,(S-Q-3184) + 12096 - 4320R)\, r^2 \\ & + ((1-R)\,(9392+2Q-2S) + 22032R - 53136)\, r \\ & + ((1-R)\,(S-Q-3184) + 12096 - 4320R)\,. \end{aligned}$$

which simplifies to

$$0 = Nr^2 - (2\,(1-R)\,(6320 + S - Q) + 31104)\, r + N$$

where $N = (1-R)(1136+S-Q)+7776$. The components of N can be written as

$$\begin{aligned}
1-R &= 1-\frac{r^4-56r^3+246r^2-56r+1}{\left(r^2-10r+1\right)^2} = \frac{\left(r^2-10r+1\right)^2-\left(r^4-56r^3+246r^2-56r+1\right)}{\left(r^2-10r+1\right)^2} \\
&= \frac{36r^3-144r^2+36r}{\left(r^2-10r+1\right)^2}
\end{aligned}$$

$$\begin{aligned}
&1136+S-Q \\
&= \left(1136+\left(\frac{\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)}{\left(r\left(1-r\right)\right)^2}\right)-\frac{\left(r^2-10r+1\right)^3}{\left(r\left(1-r\right)\right)^2}\right) \\
&= \frac{1136\left(r\left(1-r\right)\right)^2+\left(r^6-246r^5+3487r^4-10452r^3+3487r^2-246r+1\right)-\left(r^2-10r+1\right)^3}{\left(r\left(1-r\right)\right)^2} \\
&= \frac{-216r^5+4320r^4-11664r^3+4320r^2-216r}{\left(r\left(1-r\right)\right)^2}
\end{aligned}$$

so

$$\begin{aligned}
N &= \left(\left(\frac{36r^3-144r^2+36r}{\left(r^2-10r+1\right)^2}\right)\left(\frac{-216r^5+4320r^4-11664r^3+4320r^2-216r}{\left(r\left(1-r\right)\right)^2}\right)+7776\right) \\
&= \frac{\left(36r^3-144r^2+36r\right)\left(-216r^5+4320r^4-11664r^3+4320r^2-216r\right)}{\left(r^2-10r+1\right)^2\left(r\left(1-r\right)\right)^2} \\
&\quad+\frac{7776\left(r^2-10r+1\right)^2\left(r\left(1-r\right)\right)^2}{\left(r^2-10r+1\right)^2\left(r\left(1-r\right)\right)^2} \\
&= \frac{15552r^3+62208r^4+93312r^5+62208r^6+15552r^7}{\left(r^2-10r+1\right)^2\left(r\left(1-r\right)\right)^2} = \frac{15552r^3\left(r+1\right)^4}{\left(r^2-10r+1\right)^2\left(r\left(1-r\right)\right)^2} \\
N &= \frac{15552r\left(r+1\right)^4}{\left(r^2-10r+1\right)^2\left(1-r\right)^2}
\end{aligned}$$

The denominator of this expression for N is not equal to zero, because that would imply $s=0$, and we have already considered that case, and ruled it out in the derivation of the quadratic involving N. Now N could only be zero if $15552r\left(r+1\right)^4=0$, and this cannot hold because $r \neq 0$, and $r>0$ (being a ratio of probabilities) so $r \neq -1$ is ruled out. We therefore have $N \neq 0$, so the quadratic involving N can be written as $0 = r^2 - \gamma r + 1$ where $\gamma = \left(2\left(1-R\right)\left(6320 + S - Q\right) + 31104\right)/N$, which is identified because all of its components are identified.

We have now shown that $0 = r^2 - \gamma r + 1$ where γ is identified. This quadratic has solutions

$$r = \frac{1}{2}\gamma + \frac{1}{2}\sqrt{\gamma^2 - 4} \quad \text{and} \quad r = \frac{1}{\frac{1}{2}\gamma + \frac{1}{2}\sqrt{\gamma^2 - 4}}$$

so one of these must be the true value of r. Given r, we can then solve for b_0 by $b_0 = e_3^{1/3}\left(r\left(1-r\right)\right)^{1/3}$. Recall that $r = p_0/p_1$. By symmetry of the set up of the problem, if we exchanged b_0 with b_1 and exchanged p_0 with p_1 everywhere, all of the above equations would still hold. It follows that one of the above two values of r must equal p_0/p_1, and the other equals p_1/p_0. The former when substituted into $e_3\left(r\left(1-r\right)\right)$ will yield b_0^3 and the latter must by symmetry yield b_1^3. Without loss of generality imposing the constraint that $b_0 < 0 < b_1$, shows that the correct solution for r will be the one that satisfies $e_3\left(r\left(1-r\right)\right) < 0$, and so r and b_0 is identified. The remainder of the distribution of V is then given by $p_0 = r/\left(1+r\right)$, $p_1 = 1/\left(1+r\right)$, and $b_1 = -b_0 r$.

Finally, we show identification of the distribution of U. For any random variable Z, let F_Z denote the marginal cumulative distribution function of Z. By the probability mass function of the V distribution, $F_\varepsilon\left(\varepsilon\right) = \left(1-p\right)F_U\left(\varepsilon - b_1\right) + pF_U\left(\varepsilon - b_0\right)$. Letting $\varepsilon = u + \left(b_0 - b_1\right)k - b_0$ and rearranging gives

$$F_U\left(u + \left(b_0 - b_1\right)k\right) = \frac{1}{p}F_\varepsilon\left(u + b_0 + \left(b_0 - b_1\right)k\right) - \frac{1-p}{p}F_U\left(u + \left(b_0 - b_1\right)\left(k+1\right)\right)$$

so for positive integers K, $F_U\left(u\right) = R_K + \sum_{k=0}^{K-1}\left(\frac{1-p}{p}\right)^k \frac{1}{p}F_\varepsilon\left(u + b_0 + \left(b_0 - b_1\right)k\right)$ where the remainder term $R_K = r^{-K}F_U\left(u + \left(b_0 - b_1\right)K\right) \leq r^{-K}$. If $r > 1$ then $R_k \to 0$ as

$K \to \infty$, so $F_U(u)$ is identified by

$$F_U(u) = \sum_{k=0}^{\infty} \left(\frac{1-p}{p} \right)^k \frac{1}{p} F_\varepsilon \left(u + b_0 + (b_0 - b_1) k \right) \tag{13}$$

since all the terms on the right of this expression are identified, given that the distributions of ε and of V are identified. If $r < 1$, then exchange the roles of b_0 and b_1 (e.g., start by letting $\varepsilon = u + (b_1 - b_0) k - b_1$) which will correspondingly exchange p and $1-p$ to obtain $F_U(u) = \sum_{k=0}^{\infty} \left(\frac{p}{1-p} \right)^k \frac{1}{1-p} F_\varepsilon \left(u + b_1 + (b_1 - b_0) k \right)$, where now the remainder term was $R_K = r^K F_U \left(u + (b_1 - b_0) K \right) \leq r^K \to 0$ as $K \to \infty$ since now $r < 1$. The case of $r = 0$ is ruled out, since that is equivalent to $p = 1/2$.

4 Appendix B: Local GMM Asymptotic Theory

Most of the estimators in the paper are either standard GMM or well known variants of GMM. Here we summarize the application of the local GMM estimator of Lewbel (2007) to estimation based on Corollary 2.

Given the assumptions of Corollary 2, equations (2) to (9) will all hold after substituting in equation (1) and then replacing the parameters h, b_0, b_1, p, u_j, and τ_ℓ, with functions $h(X)$, $b_0(X)$, $b_1(X)$, $p(X)$, and $u_j(X)$ and replacing the unconditional expectations in these equations with conditional expectations, conditioning on $X = x$. If desired, we can further replace $b_0(X)$ and $b_1(X)$ with $g(x,0) - h(x)$ and $g(x,1) - h(x)$, respectively, to directly obtain estimates of the function $g(X, D^*)$ instead of $b_0(X)$ and $b_1(X)$.

Let $q(x)$ be the vector of all of the above listed unknown functions. Then these conditional expectations can be written as

$$E[G\left(q(x), Y\right) \mid X = x)] = 0 \tag{14}$$

for a known vector valued function G. Local GMM estimation applies generally to estimation of conditional moments in the form of equation (14)

To motivate this estimator, which is closely related to Gozalo and Linton (2000), first consider the case where all the elements of X are discrete, or more specifically, the case where X has one or more mass points and we only wish to estimate $q(x)$ at those points. Let $q_0(x)$ denote the true value of $q(x)$, and let $\theta_{x0} = q_0(x)$. If the distribution of X has a mass point with positive probability at x, then

$$E[G(\theta_x, Y) \mid X = x] = \frac{E[G(\theta_x, Y)I(X = x)]}{E[I(X = x)]}$$

so equation (14) holds if and only if $E[G(\theta_{x0}, Y)I(X = x)] = 0$. It therefore follows that under standard regularity conditions we may estimate $\theta_{x0} = q_0(x)$ using the ordinary GMM estimator

$$\widehat{\theta}_x = \arg\min_{\theta_x} \sum_{i=1}^{n} G(\theta_x, Y_i)' I(X_i = x)\Omega_n \sum_{i=1}^{n} G(\theta_x, Y_i)' I(X_i = x) \tag{15}$$

for some sequence of positive definite Ω_n. If Ω_n is a consistent estimator of $\Omega_{x0} = E[G(\theta_{x0}, Y)G(\theta_{x0}, Y)'I(X = x)]^{-1}$, then standard efficient GMM gives

$$\sqrt{n}(\widehat{\theta}_x - \theta_{x0}) \rightarrow^d N\left(0, \left[E\left(\frac{\partial G(\theta_{x0}, Y)I(X = x)}{\partial \theta_x{}'}\right)\Omega_{x0} E\left(\frac{\partial G(\theta_{x0}, Y)I(X = x)}{\partial \theta_x{}'}\right)'\right]^{-1}\right)$$

Now assume that X is continuously distributed. Then the local GMM estimator consists of applying equation (15) by replacing the average over just observations $X_i = x$ with local averaging over observations X_i in the neighborhood of x.

Assumption B1. Let X_i, Y_i, $i = 1, ..., n$, be an independently, identically distributed random sample of observations of the random vectors X, Y. The d vector X is continuously distributed with density function $f(X)$. For given point x in the interior of

$supp(X)$ having $f(x) > 0$ and a given vector valued function $G(q, y)$ where $G(q(x), y)$ is twice differentiable in the vector $q(x)$ for all $q(x)$ in some compact set $\Theta(x)$, there exists a unique $q_0(x) \in \Theta(x)$ such that $E[G(q_0(x), Y) \mid X = x] = 0$. Let Ω_n be a finite positive definite matrix for all n, as is $\Omega = plim_{n\to\infty}\Omega_n$.

Assumption B1 lists the required moment condition structure and identification for the estimator. Corollary 1 in the paper provides the conditions required for Assumption B1, in particular uniqueness of $q_0(x)$. Assumption B2 below provides conditions required for local averaging. Define $e[q(x), Y]$, $\Sigma(x)$, and $\Psi(x)$ by

$$\begin{aligned}
e[q(x), Y] &= G(q(x), Y)f(x) - E[G(q(x), Y)f(X) \mid X = x] \\
\Sigma(x) &= E\left[e(q_0(x), Y)e(q_0(x), Y)^T \mid X = x\right] \\
\Psi(x) &= E\left(\frac{\partial G[q_0(x), Y]}{\partial q_0(x)^T} f(X) \mid X = x\right)
\end{aligned}$$

Assumption B2. Let η be some constant greater than 2. Let K be a nonnegative symmetric kernel function satisfying $\int K(u)du = 1$ and $\int ||K(u)||^\eta du$ is finite. For all $q(x) \in \Theta(x)$, $E[||G(q(x), Y)f(X)||^\eta \mid X = x]$, $\Sigma(x)$, $\Psi(x)$, and $Var[[\partial G(q(x), Y)/\partial q(x)]f(X) \mid X = x]$ are finite and continuous at x and $E[G(q(x), Y)f(X) \mid X = x]$ is finite and twice continuously differentiable at x.

Define

$$S_n(q(x)) = \frac{1}{nb^d}\sum_{i=1}^{n} G[q(x), Y_i]K\left(\frac{x - X_i}{b}\right)$$

where $b = b(n)$ is a bandwidth parameter. The proposed local GMM estimator is

$$\widehat{q}(x) = \arg \inf_{q(x)\in\Theta(x)} S_n(q(x))^T \Omega_n S_n(q(x)) \tag{16}$$

The scaling of the kernal estimator $S_n(q(x))$ by b^d is convenient for deriving the properties of the estimator, but is numerically unnecessary because omitting it leaves the minimized value $\widehat{q}(x)$ unchanged.

THEOREM 3 (Lewbel 2007): Given Assumptions B1 and B2, if the bandwidth b satisfies $nb^{d+4} \to 0$ and $nb^d \to \infty$, then $\widehat{q}(x)$ is a consistent estimator of $q_0(x)$ with limiting distribution

$$(nb)^{1/2}[\widehat{q}(x)-q_0(x)] \to^d N\left[0, (\Psi(x)^T\Omega\Psi(x))^{-1}\Psi(x)^T\Omega\Sigma(x)\Omega\Psi(x)(\Psi(x)^T\Omega\Psi(x))^{-1}\int K(u)^2du\right]$$

Applying the standard two step GMM procedure, we may first estimate $\widetilde{q}(x) = \arg\inf_{q(x)\in\Theta(x)} S_n(q(x))^T S_n(q(x))$, then let Ω_n be the inverse of the sample variance of $S_n(\widetilde{q}(x))$ to get $\Omega = \Sigma(x)^{-1}$, making

$$(nb)^{1/2}[\widehat{q}(x) - q_0(x)] \to^d N\left[0, (\Psi(x)^T\Omega\Psi(x))^{-1}\int K(u)^2du\right]$$

where $\Psi(x)$ can be estimated using

$$\Psi_n(x) = \frac{1}{nb^d}\sum_{i=1}^{n}\frac{\partial G[\widehat{q}(x), Y_i]}{\partial\widehat{q}(x)^T}K\left(\frac{x - X_i}{b}\right)$$

At the expense of some additional notation, the two estimators (15) and (16) can be combined to handle X containing both discrete and continuous elements, by replacing the kernel function in S_n with the product of a kernel over the continuous elements and an indicator function for the discrete elements, as in Li and Racine (2003).

References

[1] Bianchi, M. (1997), ”Testing for Convergence: Evidence from Nonparametric Multimodality Tests,” Journal of Applied Econometrics, 12, 393-409.

[2] Gozalo, P, and Linton, O. (2000). Local Nonlinear Least Squares: Using Parametric Information in Non-parametric Regression. Journal of econometrics, 99, 63-106.

[3] Lewbel, A. (2007) “A Local Generalized Method of Moments Estimator,” Economics Letters, 94, 124-128.

[4] Li, Q. and J. Racine (2003), ”Nonparametric estimation of distributions with categorical and continuous data,” Journal of Multivariate Analysis, 86, 266-292

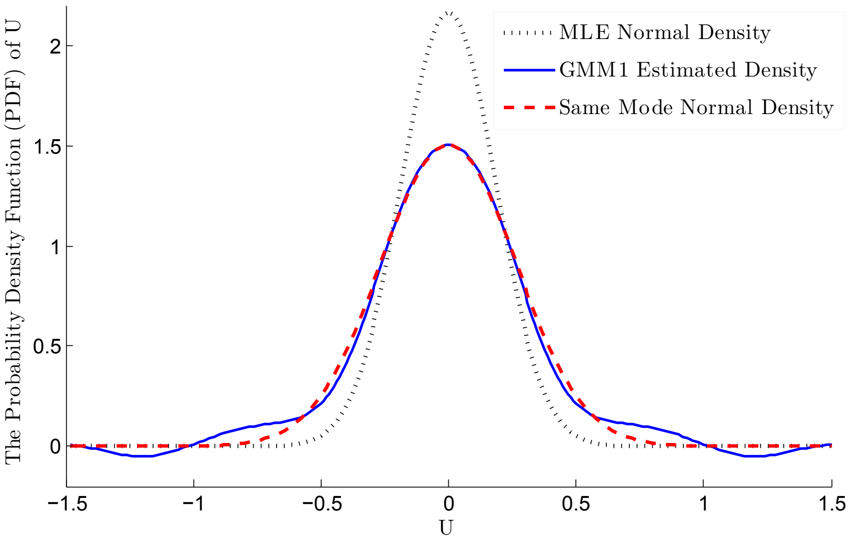

Figure 1: The estimated probability density function of U, using 1970 share data