MEASUREMENT ERROR MODELS

XIAOHONG CHEN and HAN HONG and DENIS NEKIPELOV[1]

Key words: Linear or nonlinear errors-in-variables models, classical or nonclassical measurement errors, attenuation bias, instrumental variables, double measurements, deconvolution, auxiliary sample
JEL Classification: C1, C3

1 Introduction

Many economic data sets are contaminated by the mismeasured variables. The problem of measurement errors is one of the most fundamental problems in empirical economics. The presence of measurement errors causes biased and inconsistent parameter estimates and leads to erroneous conclusions to various degrees in economic analysis. Techniques for addressing measurement error problems can be classified along two dimensions. Different techniques are employed in linear errors-in-variables (EIV) models and in nonlinear EIV models. (In this article, a "linear" EIV model means it is linear in both the mismeasured variables and the parameters of interest; a "nonlinear" EIV model means it is nonlinear in the mismeasured variables.) Different methods are used to treat classical measurement errors and nonclassical measurement errors. (A measurement error is "classical" if it is independent of the latent true variable; otherwise it is "nonclassical".) Since various methods for linear EIV models with classical measurement errors are already known and are widely applied in empirical economics, in this survey we shall focus more on recent theoretical advances on methods for identification and estimation of nonlinear EIV models with classical or nonclassical measurement errors. While measurement error problems can be as severe with time series data as with cross sectional data, in this survey we shall focus on cross

[1] Department of Economics, New York University and Department of Economics, Stanford University and Department of Economics, Duke University, USA. The authors acknowledge generous research supports from the NSF (Chen and Hong) and the Sloan Foundation (Hong). This is an article prepared for the Journal of Economic Literature. The authors thank the editor Roger Gordon for suggestions and Shouyue Yu for research assistance. The usual disclaimer applies.

sectional data and maintain the assumption that the data are independently and identically distributed.

Due to the importance of the measurement error problems, there are huge amount of papers and several books on measurement errors; hence it is impossible for us to review all the existing literature. Instead of attempting to cover as many papers as we could, we intend to survey relatively recent developments in econometrics and statistics literature on measurement error problems. Reviews of earlier results on this subject can also be found in Fuller (1987), Carroll, Ruppert, and Stefanski (1995), Wansbeek and Meijer (2000), Bound, Brown, and Mathiowetz (2001), Hausman (Autumn, 2001) and Moffit and Ridder (to appear), to name only a few.

In this survey we aim at introducing recent theoretical advances in measurement errors to applied researchers. Instead of stating technical conditions rigorously, we mainly describe key ideas for identification and estimation, and refer readers to the original papers for technical details. Since most of the theoretical results on nonlinear EIV models are very recent, there are not many empirical applications yet. We shall mention applications of these new methods whenever they are currently available. The rest of the survey is organized as follows. Section 2 briefly mentions results for linear EIV models with classical measurement errors. Section 3 reviews results on nonlinear EIV models with classical measurement errors. Section 4 presents very recent results on nonlinear EIV models with nonclassical measurement errors, including misclassification in models with discrete variables. Section 5 reviews results on bounds for parameters of interest when the EIV models are only partially identified under weak assumptions. Section 6 briefly concludes.

2 Linear EIV Model With Classical Errors

The classical measurement error assumption maintains that the measurement errors in any of the variables in the data set are independent of all the true variables that are the objects of interest. The implication of this assumption in the linear least square regression model $y_i^* = x_i^{*'}\beta + \epsilon_i$ is well understood and is usually described in a standard econometrics textbook. Under this assumption, measurement errors in the dependent variable $y_i = y_i^* + v_i$ do not lead to inconsistent estimate of the regression coefficients, as can be seen by rewriting the model in y_i:

$$y_i = x_i^{*'} \beta + \epsilon_i + v_i = x_i^{*'} \beta + \omega_i$$

The only consequence of the presence of measurement errors in the dependent variables is that they inflate the standard errors of these regression coefficient estimates. On the other hand, independent errors that are present in the observations of the regressors $x_i = x_i^* + \eta_i$ lead to attenuation bias in a simple univariate regression model and to inconsistent regression coefficient estimates in general.

Attenuation bias: Consider a univariate classical linear regression model

$$y = \alpha + \beta x^* + \epsilon, \ E(x^* \epsilon) = 0, \tag{1}$$

where x^* can only be observed with an additive, independent measurement error $\eta \sim \left(0, \sigma_\eta^2\right)$:

$$x = x^* + \eta. \tag{2}$$

Then, the regression of y on x can be obtained by inserting (2) into (1):

$$y = \alpha + \beta x + u, \ u = \epsilon - \beta \eta. \tag{3}$$

Given a random sample of n observations (y_i, x_i) on (y, x), the least squares estimator is given by:

$$\hat{\beta} = \frac{\sum_{j=1}^n \left(x_j - \bar{x}\right) y_j}{\sum_{j=1}^n \left(x_j - \bar{x}_j\right)^2}. \tag{4}$$

Since x and u are correlated with each other

$$Cov[x, u] = Cov[x^* + \eta, \epsilon - \beta \eta] = -\beta \sigma_\eta^2 \neq 0,$$

the least squares estimator should be inconsistent. Its probability limit is:

$$\text{plim} \hat{\beta} = \beta + \frac{Cov\left(x, u\right)}{Var\left(x\right)} = \beta - \frac{\beta \sigma_\eta^2}{\sigma_*^2 + \sigma_\eta^2} = \beta \frac{\sigma_*^2}{\sigma_*^2 + \sigma_\eta^2}, \tag{5}$$

where $\sigma_*^2 = Var\left(x^*\right)$. Since σ_η^2 and σ_*^2 are both positive, $\hat{\beta}$ is inconsistent for β with an attenuation bias. This result can easily be extended to a multivariate linear regression model. In the multivariate case, one should notice that even if only the measurement on a single regressor is error-prone, the coefficients on all regressors are generally biased.

The importance of measurement errors in analyzing the empirical implications of economic theories is highlighted in Milton Friedman's seminal book on the consumption theory of permanent income hypothesis (Friedman (1957)). In Friedman's model, both consumption and income are composed of a permanent component and a transitory component that can be due to measurement errors or genuine fluctuations. The marginal propensity to consume relates the permanent component of consumption to the permanent income component. Friedman shows that because of the attenuation bias, the slope coefficient of a regression of observed consumption on observed income would lead to an underestimate of the marginal propensity to consume.

Frisch bounds Econometric work on linear models with classical independent additive measurement error dates back to **Fricsh (1934)**, who derives the bounds on the slope and the constant term by least squares estimation in different directions. Consider a univariate linear regression model with measurement errors defined in (1) to (3). In addition to the bias in the slope coefficient presented above, the estimate of the intercept is given by

$$\hat{\alpha} = \bar{y} - \hat{\beta} \times \bar{x}, \tag{6}$$

and has a probability limit given by

$$\text{plim}\hat{\alpha} = E[\alpha + \beta x^* + \epsilon] - \frac{\beta\sigma_*^2}{\sigma_*^2 + \sigma_u^2} E[x^* + \eta] = \alpha + \frac{\beta\sigma_u^2}{\sigma_*^2 + \sigma_u^2}\mu^*,$$

where $\mu^* = Ex^*$.

Consider running a regression in the opposite direction in the second step. Rewrite the regression model (3) as

$$x = -\frac{\alpha}{\beta} + \frac{1}{\beta} y - \frac{\epsilon - \beta\eta}{\beta}. \tag{7}$$

The inverse regression coefficient and intercept estimates are defined by:

$$\hat{\beta}_{rev} = \frac{1}{b_{rev}} \quad \text{where} \quad b_{rev} = \frac{\sum_{i=1}^n (x_i - \bar{x})\, y_i}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \text{and} \quad \hat{\alpha}_{rev} = \bar{y} - \hat{\beta}_{rev}\bar{x}. \tag{8}$$

The probability limits of these slope and constant terms can be derived following the same procedure as above:

$$\text{plim}\hat{\beta}_{rev} = \text{plim}\frac{1}{b_{rev}} = \frac{Var\,(y)}{Cov\,(x, y)} = \frac{\beta^2\sigma_*^2 + \sigma_\epsilon^2}{\beta\sigma_*^2} = \beta + \frac{\sigma_\epsilon^2}{\beta\sigma_*^2}, \tag{9}$$

and

$$\text{plim}\hat{\alpha}_{rev} = \alpha + \beta\mu^* - \left(\beta + \frac{\sigma_\epsilon^2}{\beta\sigma_*^2}\right)\mu^* = \alpha - \frac{\mu^*\sigma_\epsilon^2}{\beta\sigma_*^2}. \tag{10}$$

Clearly, the "true" coefficients α and β lie in the bounds formed by the probability limits of the direct estimators in (4) and (6) and the reverse estimators in (8).

Measurement error models can be regarded as a special case of models with endogenous regressors; hence the method of **Instrumental Variables (IV)** is a popular approach to obtaining identification and consistent point estimates of parameters of interest in linear regression models with classical independent additive measurement errors. For example, assuming there is an IV w such that $E(wx) \neq 0$ and $E(wu) = 0$ for the model (3), then the standard instrumental variable estimator of β will be consistent. In addition, one can apply Hausman test to check the presence of classical measurement errors in linear regression models. In practice, a valid IV often comes from a second measurement of the error-prone true variable: $w_i = x_i^* + v_i$, which is subject to another independent measurement error v_i. Because w_i is mean independent of (ϵ_i, η_i) but is correlated with the first measurement $x_i = x_i^* + \eta_i$, the second measurement w_i is a valid IV for the regressor x_i in the linear regression model (3): $y_i = \alpha + \beta x_i + u_i$, $u_i = \epsilon_i - \beta\eta_i$.

3 Nonlinear EIV Model With Classical Errors

It is well known that, without additional information or functional form restrictions, a general nonlinear EIV model cannot be identified. As shown in Amemiya (1985), standard IV assumption (i.e., mean correlated with mismeasured regressor and mean uncorrelated with regression error) that allows for point identification of linear EIV regression models is no longer sufficient for identification of nonlinear EIV regression models, even when the measurement error is additively independent of the latent true regressor. In section 5 we discuss some results on partial identification and bound analysis of nonlinear EIV models, under weak assumptions. In this and the next sections we focus on point identification under various additional restrictions, assuming either known or parametric distributions of measurement errors, or double measurements of mismeasured regressors, or strong notions of instrumental variables, or auxiliary samples.

3.1 Nonlinear EIV models via Deconvolution

Almost all the methods for identification of nonlinear EIV models with classical measurement errors are various extensions of the method of **deconvolution**. Consider a general nonlinear EIV moment restriction model

$$Em\left(y^*;\beta\right)=0$$

under the classical measurement error assumption: $y_i = y_i^* + \varepsilon_i$, where only $y_i \in \mathbb{R}^k$ is observed. Here for simplicity we do not distinguish between dependent and independent variables and use y^* to denote the entire vector of true unobserved variables. Suppose one knows the characteristic function $\phi_\varepsilon\left(t\right) = Ee^{it\varepsilon_i}$ of the classical measurement errors ε_i. Given that the measurement error is independent from the latent variables y_i^*, the characteristic function of y_i^* can be recovered from the ratio of the characteristic functions $\phi_y(t)$ and $\phi_\varepsilon\left(t\right)$ of y_i and ε_i:

$$\phi_{y^*}\left(t\right)=\phi_y(t)/\phi_\varepsilon\left(t\right),$$

where an estimate $\hat{\phi}_y\left(t\right)$ of $\phi_y\left(t\right)$ can be obtained using a smooth version of $\frac{1}{n}\sum_{i=1}^n e^{ity_i}$. Then $\phi_{y^*}\left(t\right)$ can be estimated by

$$\hat{\phi}_{y^*}\left(t\right)=\hat{\phi}_y(t)/\phi_\varepsilon\left(t\right).$$

Once the characteristic function of y^* is known, its density can be recovered from by the inverse Fourier transformation of the corresponding characteristic function

$$\hat{f}\left(y^*\right)=\left(\frac{1}{2\pi}\right)^k\int\hat{\phi}_{\mathbf{y}^*}\left(\mathbf{t}\right)e^{-iy^{*\prime}\mathbf{t}}d\mathbf{t}.$$

For each β, a sample analog of the moment condition $Em\left(y^*;\beta\right)=0$ can then be estimated by

$$\int m\left(y^*;\beta\right)\hat{f}\left(y^*\right)dy^*.$$

One can obtain a semiparametric Generalized Method of Moment (GMM) estimator as a minimizer over β of an Euclidean distance of the above estimated system of moments from zeros.

There are many papers in statistic literature on estimation of nonparametric or semiparametric EIV models using the deconvolution method, assuming completely known distributions of the classical measurement errors. See e.g., Carroll and Hall (1988), Fan (1991) and Fan and Truong (1993) for the optimal convergence rates for nonparametric deconvolution problems; Taupin (2001) and Butucea and Taupin (2005) for semiparametric estimation.

The original deconvolution method assumes that the distribution of the classical measurement error is completely known, and it is later extended to allow for parametrically specified measurement error distribution, or double measurements, or other strong notions of instrumental variables. We shall discuss these extensions subsequently.

3.2 Nonlinear models with parametric measurement error distributions

For certain parametric families of the measurement error distribution the characteristic function of the measurement error $\phi_\varepsilon(t)$ can be parameterized and its parameters can be estimated jointly with the parameter of the econometric model β. **Hong and Tamer (2003)** assume that the marginal distributions of the measurement errors are Laplace (double exponential) with zero means and unknown variances and the measurement errors are independent of the latent variables and are independent of each other. Under these assumptions, they derive simple revised moment conditions in terms of the observed variables that lead to a simple estimator for the case of nonlinear moment models under the assumption that the measurement error is classical (so that it is independent and and additively separable from the latent regressor) when no data on additional measurements are available.

When the distributions of ε are independent Laplace (double exponential), its characteristic function takes the form of

$$\phi_\varepsilon(t) = \prod_{j=1}^{k} \left(1 + \frac{1}{2}\sigma_j^2 t_j^2\right)^{-1}.$$

Using this characteristic function, Hong and Tamer (2003) show that the moment condition for the latent random vector $\mathbf{y}^*$ expressed as $Em(\mathbf{y}^*;\beta) = 0$ can be translated into the moment condition for the observable random variable $\mathbf{y}$ as

$$Em(\mathbf{y}^*;\beta) = Em(\mathbf{y};\beta) + \sum_{l=1}^{k}\left(-\frac{1}{2}\right)^{l} \sum \cdots \sum_{j_1<\cdots<j_l} \sigma_{j1}^2 \cdots \sigma_{jl}^2 E\frac{\partial^{2l}}{\partial y_{j1}^2 \cdots \partial y_{jl}^2} m(\mathbf{y};\beta).$$

Consider the following model as an example:

$$E\left[y \mid x^*\right] = g(x^*;\beta),\ x = x^* + \varepsilon,$$

where $g(\cdot;\cdot)$ is a known twice differentiable function and x^* is a latent variable defined on $\mathbb{R}$ such that the conditional variance $Var(y|x^*)$ is finite. This model implies the unconditional moment restriction,

$$E\left[\mathbf{h}(x^*)(y - g(x^*;\beta))\right] = 0$$

for a $h \times 1$ ($h > dim(\beta)$) vector of measurable functions $\mathbf{h}(\cdot)$. Then, the revised moment conditions in terms of observed variables are

$$\begin{aligned} E\Bigg[\mathbf{h}(x)(y - g(x;\beta)) - \frac{1}{2}\sigma^2(\mathbf{h}^{(2)}(x)y - \mathbf{h}^{(2)}(x)g(x;\beta) \\ - 2\mathbf{h}^{(1)}(x)g^{(1)}(x;\beta) - \mathbf{h}(x)g^{(2)}(x;\beta))\Bigg] = 0. \end{aligned}$$

For each candidate parameter value β, the right hand side of the revised moment conditions can be estimated from the sample analog by replacing the expectation with the empirical sum. Define the moment function

$$m\left(\mathbf{y};\beta,\,\sigma\right) = m\left(\mathbf{y};\beta\right) + \sum_{l=1}^{k}\left(-\frac{1}{2}\right)^{l}\sum_{j_1<\cdots<j_l}\!\!\cdots\!\sum \sigma_{j1}^2\cdots\sigma_{jl}^2\frac{\partial^{2l}}{\partial y_{j1}^2\cdots\partial y_{jl}^2}m\left(\mathbf{y};\beta\right).$$

The revised moment condition:

$$E\,m\left(\mathbf{y};\beta,\,\sigma\right) = 0$$

can be used to obtain point estimates of both the parameter of the econometric model β and the parameters characterizing the distribution of the measurement error $\sigma \equiv \{\sigma_j, j = 1, ..., k\}$, provided that this revised moment condition is sufficient to point identify both sets of parameters. Explicitly, for some symmetric positive definite $h \times h$ weighting matrix W_n, the GMM estimators for β and σ (identified via the revised moment condition) are given by:

$$(\hat{\beta},\hat{\sigma}) = \underset{\beta,\sigma}{\operatorname{argmin}}\left(\frac{1}{n}\sum_{i=1}^{n}m\left(\mathbf{y_i};\beta,\sigma\right)\right)' W_n\left(\frac{1}{n}\sum_{i=1}^{n}m\left(\mathbf{y_i};\beta,\sigma\right)\right).$$

Hong and Tamer (2003) further prove the consistency and asymptotic normality of the revised method of moment estimator under the assumption of global point identification and other regularity conditions (including the compactness of the parameter set, uniform boundedness of the moments of the model, Laplacian characteristic function for the distribution of the observation errors, and Lipschitz condition for the partial derivative of the system of moments with respect to the parameter vector). More precisely, under some regularity assumptions they establish:

$$(\hat{\beta},\hat{\sigma}) \xrightarrow{p} (\beta,\sigma);\ \sqrt{n}\left((\hat{\beta},\hat{\sigma})-(\beta,\sigma)\right) \xrightarrow{d} N\left(0,(A'WA)^{-1}(A'W\Omega WA)(A'WA)^{-1}\right),$$

where $A \equiv E\frac{\partial}{\partial(\beta,\sigma)}m\left(\mathbf{y};\beta,\sigma\right)$, $W = \text{plim}W_n$, and $\Omega = Em\left(\mathbf{y};\beta,\sigma\right)m\left(\mathbf{y};\beta,\sigma\right)'$. The authors also provide the result for the two-step GMM that uses the estimate of the weighting matrix

$$\hat{W}_n = \left(\frac{1}{n}\sum_{i=1}^{n} m\left(\mathbf{y};\hat{\beta},\hat{\sigma}\right)m\left(\mathbf{y};\hat{\beta},\hat{\sigma}\right)'\right)^{-1}$$

obtained from the first step.

Even if the revised moment condition $E[m\left(\mathbf{y};\beta,\,\sigma\right)] = 0$ cannot point identify the parameter β, it still contains useful information about β that can be exploited using the information about $\sigma_1^2,\ldots,\sigma_k^2$. In this case, under certain conditions we can provide the bounds for the parameter β giving partial identification information. We know that for any sensible inference for all $j = 1,\ldots,k$ the variance of the measurement errors should be smaller than the variance of the "signal"

$$0 \leq \sigma_j^2 \leq \sigma_{y_j}^2, \tag{11}$$

where $\sigma_{y_j}^2$ is the variance of the observed random variable y_j. Then, the set of observationally equivalent parameter values can be defined as:

$$V = \left\{\mathbf{b}\in\mathcal{B}|\,\eta_0(\mathbf{b}) \leq 0 \leq \eta_1(\mathbf{b})\right\},$$

where

$$\eta_0(\mathbf{b}) = Em\left(\mathbf{y};\mathbf{b}\right) + \textstyle\sum_{l=1}^{k}\sum\limits_{j_1<\cdots<j_l}\sigma_{y_{j_1}}^2\cdots\sigma_{y_{j_l}}^2\left[\left(-\frac{1}{2}\right)^l E\frac{\partial^{2l}}{\partial y_{j_1}^2\cdots\partial y_{j_l}^2}m\left(\mathbf{y};\mathbf{b}\right)\right]^-,$$

$$\eta_1(\mathbf{b}) = Em\left(\mathbf{y};\mathbf{b}\right) + \textstyle\sum_{l=1}^{k} \sum\limits_{j_1<\cdots<j_l} \sigma^2_{y_{j_1}} \cdots \sigma^2_{y_{j_l}} \left[\left(-\frac{1}{2}\right)^l E \frac{\partial^{2l}}{\partial y^2_{j_1}\cdots\partial y^2_{j_l}} m\left(\mathbf{y};\mathbf{b}\right) \right]^+ .$$

Based on this, the identified features of the model can be estimated by a Modified Method of Moments (MMM) estimator. Define the moment objective as a sum of the weighted modified moment criteria

$$\begin{aligned} T(\mathbf{b}) = &\left\{\left[\eta_0(\mathbf{b})\mathbf{1}\left(\eta_0(\mathbf{b})>0\right)\right]' \mathbf{W} \left[\eta_0(\mathbf{b})\mathbf{1}\left(\eta_0(\mathbf{b})>0\right)\right]\right\} \\ &+ \left\{\left[\eta_1(\mathbf{b})\mathbf{1}\left(\eta_1(\mathbf{b})<0\right)\right]' \mathbf{W} \left[\eta_1(\mathbf{b})\mathbf{1}\left(\eta_1(\mathbf{b})<0\right)\right]\right\} \end{aligned}$$

and its sample analog:

$$\begin{aligned} Q_n(\mathbf{b}) = &\left\{\left[\eta_{0n}(\mathbf{b})1\left(\eta_{0n}(\mathbf{b})>0\right)\right]' \mathbf{W} \left[\eta_{0n}(\mathbf{b})1\left(\eta_{0n}(\mathbf{b})>0\right)\right]\right\} \\ &+ \left\{\left[\eta_{1n}(\mathbf{b})1\left(\eta_{1n}(\mathbf{b})<0\right)\right]' \mathbf{W} \left[\eta_{1n}(\mathbf{b})1\left(\eta_{1n}(\mathbf{b})<0\right)\right]\right\}, \end{aligned} \tag{12}$$

where we use the sample analogs of the corresponding moment equations

$$\eta_{0n}(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n} m\left(\mathbf{y_i};\mathbf{b}\right) + \sum_{l=1}^{k} \sum_{j_1<\cdots<j_l} \sigma^2_{y_{j_1}n} \cdots \sigma^2_{y_{j_l}n} \left[\left(-\frac{1}{2}\right)^l \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^{2l}}{\partial y^2_{j_1}\cdots\partial y^2_{j_l}} m\left(\mathbf{y}_i;\mathbf{b}\right) \right]^-$$

and

$$\eta_{1n}\left(\mathbf{b}\right) = \frac{1}{n}\sum_{i=1}^{n} m\left(\mathbf{y_i};\mathbf{b}\right) + \sum_{l=1}^{k} \sum_{j_1<\cdots<j_l} \sigma^2_{y_{j_1}n} \cdots \sigma^2_{y_{j_l}n} \left[\left(-\frac{1}{2}\right)^l \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^{2l}}{\partial y^2_{j_1}\cdots\partial y^2_{j_l}} m\left(\mathbf{y}_i;\mathbf{b}\right) \right]^+$$

and

$$\sigma^2_{y_j n} = \frac{1}{n}\sum_{i=1}^{n} \left(y_{i,j} - \frac{1}{n}\sum_{i'=1}^{n} y_{i',j} \right)^2 .$$

Then, the consistent MMM estimator is giving the set of possible values of the parameter of the econometric model in the form:

$$V_n = \left\{ \mathbf{b} \in \mathcal{B} \mid \quad Q_n(\mathbf{b}) \leq \underset{\mathbf{c}\in\mathcal{B}}{argmin}\ Q_n(\mathbf{c}) + \gamma_n \right\}$$

where $\gamma_n > 0$ and $\gamma_n \to 0$ as $n \to \infty$. The assumption on the distribution of the measurement errors seems to be very strong. However, Hong and Tamer (2003) show that the estimation is robust for a wide variety of specifications of the measurement error distribution.

3.3 Nonlinear EIV models with double measurements

3.3.1 Models nonlinear-in-variables but linear-in-parameters

The double measurement instrumental variable method for linear regression models has been generalized by **Hausman, Newey, Ichimura, and Powell (1991)** to polynomial regression models in which the regressors are polynomial functions of the error-prone variables. The following is a simplified version of the polynomial regression model that they considered:

$$y = \sum_{j=0}^{K} \beta_j (x^*)^j + r'\phi + \epsilon.$$

Among the two sets of regressors x^* and r, r is precisely observed but x^* is only observed with classical errors. In particular, two measurements of x^*, x and w, are observed which satisfy

$$x = x^* + \eta \quad \text{and} \quad w = x^* + v.$$

We will focus on identification of population moments. For convenience, assume that ϵ, η and v are mutually independent and they are independent of all the true regressors in the model.

First assume that $\phi = 0$, then identification of β depends on population moments $\xi_j \equiv E\left(y(x^*)^j\right), j = 0, \ldots, K$ and $\zeta_m \equiv E(x^*)^m, m = 0, \ldots, 2K$, which are the elements of the population normal equations for solving for β. Except for ξ_0 and ζ_0, these moments depend on x^* which is not observed, but they can be solved from the moments of observable variables Exw^{j-1}, Ew^l for $j = 0, \ldots, 2K$ and $Eyw^j, j = 0, \ldots, K$. Define $\nu_k = Ev^k$. Then the observable moments satisfy the following relations:

$$\begin{aligned} Exw^j =& E\left(x^* + \eta\right)\left(x^* + v\right)^j = E \sum_{l=0}^{j} \binom{j}{l} \left(x^* + \eta\right)(x^*)^l v^{j-l} \\ =& \sum_{l=0}^{j} \binom{j}{l} \zeta_{l+1} \nu_{j-l}, \quad j = 1, 2K-1, \end{aligned} \tag{13}$$

and

$$Ew^j = E\left(x^* + v\right)^j = E\sum_{l=0}^{j}\binom{j}{l}(x^*)^l v^{j-l} = \sum_{l=0}^{j}\binom{j}{l}\zeta_l\nu_{j-l}, \quad j = 1,\ldots,2K, \tag{14}$$

and

$$Eyw^j = Ey\left(x^* + v\right)^j = E\sum_{l=0}^{j}\binom{j}{l}y(x^*)^l v^{j-l} = \sum_{l=0}^{j}\binom{j}{l}\xi_l\nu_{j-l}, \quad j = 1,\ldots,K. \tag{15}$$

Since $\nu_1 = 0$, we have a total of $(5K-1)$ unknowns in $\zeta_1,\ldots,\zeta_{2K}$, $\xi_1,\ldots,\xi_K$ and $\nu_2,\ldots,\nu_{2K}$. Equations (13), (14) and (15) give a total of $5K-1$ equations that can be used to solve for these $5K-1$ unknowns. In particular, the $4K-1$ equations in (13) and (14) jointly solve for $\zeta_1,\ldots,\zeta_{2K},\nu_2,\ldots,\nu_{2K}$. Subsequently, given knowledge of these ζ's and ν's, ξ's can then be recovered from equation (15). Finally, we can use these identified quantities of $\xi_j, j = 0,\ldots,K$ and $\zeta_m, m = 0,\ldots,2K$ to recover the parameters β from the normal equations

$$\xi_l = \sum_{j=0}^{K}\beta_j\zeta_{j+l}, \quad l = 0,\ldots,K.$$

When $\phi \neq 0$, Hausman, Newey, Ichimura, and Powell (1991) note that the normal equations for the identification of β and ϕ depends on a second set of moments Eyr, Err' and $Er(x^*)^j, j = 0,\ldots,K$, in addition to the first set of moments $\xi's$ and $\zeta's$. Since Eyr and Err' can be directly observed from the data, it only remains to identify $Er(x^*)^j, j = 0,\ldots,K$. But these can be solved from the following system of equations, for $j = 0,\ldots,K$:

$$Erw^j = Er\left(x^* + v\right)^j = E\sum_{l=0}^{j}\binom{j}{l}r(x^*)^l v^{j-l} = \sum_{l=0}^{j}\binom{j}{l}\left(Er(x^*)^l\right)\nu_{j-l}, j = 0,\ldots,K.$$

In particular, using the previously determined ν coefficients, the jth row of the previous equation can be solved recursively to obtain

$$Er(x^*)^j = Erw^j - \sum_{l=0}^{j-1}\binom{j}{l}\left(E(x^*)^l r\right)\nu_{j-l}.$$

Once all these elements of the normal equations are identified, the coefficients β and ϕ can then be solved from the normal equations $[EyZ', \ Eyr]' = D\,[\beta', \ \phi']'$, where $Z = \left(1, (x^*), \ldots, (x^*)^K\right)'$ and $D = E\left[(Z'r')', \ (Z'r')\right]$.

Hausman, Newey, and Powell (1995) apply the identification and estimation methods proposed in Hausman, Newey, Ichimura, and Powell (1991) to estimation of Engel curve specified in the Gorman form using 1982 Consumer Expenditure Survey (CEX) data set.

3.3.2 General nonlinear models with double measurements

Often times the characteristic function of the measurement errors $\phi_\varepsilon\left(t\right)$ might not be known. However, if two independent measurements of the latent true variable y^* with additive errors are observed and the errors are i.i.d, an estimate of $\hat{\phi}_\varepsilon\left(t\right)$ can be obtained using the two independent measurements.

Li (2002) provides one method to do this. In particular, Li (2002) adopts the characteristic function approach to the estimation of nonlinear models with classical measurement errors, without assuming functional forms of the measurement error distributions. Suppose the dependent variable y is determined by the unobservable independent random vector x^* and a random disturbance u through a nonlinear relationship $y = g(\mathbf{x}^*; \beta) + u$, where the random disturbance u is independent from the vector x^* with $Eu = 0$, $E(u^2) = \sigma_0^2$, and $\mathbf{x}^* = \left(x^{(*1)}, \ldots, x^{(*K)}\right) \in \mathbb{R}^K$ is the unobservable random vector. Li (2002) assumes that two proxies $\mathbf{z}_l$, $l = 1, 2$ for $\mathbf{x}^*$ are observed:

$$\mathbf{z}_l = \mathbf{x}^* + \varepsilon_l, \quad E(\varepsilon_l) = 0, \quad l = 1, 2$$

with individual elements $z_l^{(k)}, k = 1, \ldots, K$ and $\varepsilon_l^{(k)}, k = 1, \ldots, K$. The measurement errors $(\varepsilon_l, l = 1, 2)$ and the unobservable vector of regressors $\mathbf{x}^*$ are mutually independent. In addition, $(\varepsilon_l, l = 1, 2)$ are independent of u conditional on the latent regressors $\mathbf{x}^*$. In fact, one only needs u to be mean independent of $\mathbf{x}^*$ and ε_l: $E\left(u|\mathbf{x}^*, \varepsilon_l\right) = 0$. Furthermore, Li (2002) assumes that the characteristic functions of the components of the latent regressor $\mathbf{x}^*$ and the measurement errors ε are not equal to zero in the entire space. This assumption allows the author to identify the measurement errors by restricting their distributions from decaying ”too fast” at the infinity.

The assumption about the mean independence of random disturbance u from the latent regressor $\mathbf{x}^*$ implies that the conditional expectation of the dependent variable y given the knowledge of the latent vector $\mathbf{x}^*$ is determined solely by the function $g(\cdot)$, i.e. $E\left(y|x^*\right) = g\left(\mathbf{x}^*, \beta\right)$. From this expression we can obtain the expressions for the conditional expectation of the dependent variable given the observable proxies for $\mathbf{x}^*$ and the conditional distribution of the latent variable given the proxy variable (which is determined by the distribution of the classical measurement error). In particular, for two observable proxy variables $l = 1, 2$,

$$\begin{aligned} E\left(y|\mathbf{z}_l\right) =& E\left[E\left(y|\mathbf{x}^*, \mathbf{z}_l\right)|\mathbf{z}_l\right] = E\left[E\left(y|\mathbf{x}^*, \varepsilon_l\right)|\mathbf{z}_l\right] \\ =& E\left[g\left(\mathbf{x}^*; \beta\right)|\mathbf{z}_l\right] = \int g\left(\mathbf{x}^*; \beta\right) f_{\mathbf{x}^*|\mathbf{z}_l}\left(\mathbf{x}^*|\mathbf{z}_l\right) d\mathbf{x}^*. \end{aligned}$$

In the above, the third equality follows from $\epsilon_l \perp u|\mathbf{x}^*$. Therefore if one can obtain a nonparametric estimate $\hat{f}_{\mathbf{x}^*|\mathbf{z}_l}\left(\mathbf{x}^*|\mathbf{z}_l\right)$ of the conditional distribution of the latent variable given the observable proxy variable $f_{\mathbf{x}^*|\mathbf{z}_l}\left(\mathbf{x}^*|\mathbf{z}_l\right)$, then one can run a nonlinear regression of y on

$$\int g\left(\mathbf{x}^*; \beta\right) \hat{f}_{\mathbf{x}^*|\mathbf{z}_l}\left(\mathbf{x}^*|\mathbf{z}_l\right) d\mathbf{x}^*$$

to obtain an consistent estimate of β.

The previous discussion suggests that the independence of the latent variable given the measurement error and the additive structure of the dependence between the proxy variable and the latent variable allows one to obtain the expression for the characteristic function of the measurement error from the characteristic function of the latent variable and the characteristic function of the observable proxy variable. In the case when separate measurements are available we can avoid the need for the unknown distribution of the latent variable in this procedure. To identify the conditional distribution of the latent variable given proxy $f_{\mathbf{x}^*|\mathbf{z}_l}\left(\mathbf{x}^*|\mathbf{z}_l\right)$, Li (2002) starts by showing that under the imposed assumptions about the distributions and the characteristic functions of the latent variables, random disturbances and the measurement errors, probability density functions of $x^{(*k)}$ and $\varepsilon_l^{(k)}, l = 1, 2$ can be uniquely determined from the joint distribution of $(z_1^{(k)}, z_2^{(k)})$. The joint characteristic function of the proxy variables $(z_1^{(k)}, z_2^{(k)})$ can be obtained by definition as

$$\psi_k\left(u_1, u_2\right) = Ee^{iu_1 z_1^{(k)} + iu_2 z_2^{(k)}}.$$

Then the characteristic functions for the components of the latent vector and the measurement errors $x^{(*k)}, \varepsilon_1^{(k)}$, and $\varepsilon_2^{(k)}$, denoted $\phi_x^{(*k)}(t)$, $\phi_{\varepsilon_1}^{(k)}(t)$ and $\phi_{\varepsilon_2}^{(k)}(t)$, can be derived from $\psi_k\left(u_1, u_2\right)$ through the relations:

$$\begin{aligned} \phi_x^{(*k)}(t) =& \exp\left\{\int_0^t \frac{\partial\psi_k(0,u_2)/\partial u_1}{\psi_k(0,u_2)} du_2\right\}, \\ \phi_{\varepsilon_1}^{(k)}(t) =& \frac{\psi_k(t,0)}{\phi_x^{(*k)}(t)}, \\ \phi_{\varepsilon_2}^{(k)}(t) =& \frac{\psi_k(0,t)}{\phi_x^{(*k)}(t)}. \end{aligned} \tag{16}$$

The expressions (16) are obtained using the independence and separability assumptions. To derive these expressions, note first that due to the additive separability $z_l^{(k)} = x^{(*k)} + \varepsilon_l^{(k)}$, so that substitution into the expression for the characteristic function of the proxy variables gives

$$\psi_k\left(u_1, u_2\right) = Ee^{iu_1\left(x^{(*k)}+\varepsilon_1^{(k)}\right)+iu_2\left(x^{(*k)}+\varepsilon_2^{(k)}\right)}.$$

The independence of $\varepsilon_1^{(k)}$ from $x^{(*k)}$ and $\varepsilon_2^{(k)}$ implies that

$$E\left(\varepsilon_1^{(k)}|x^{(*k)}, \varepsilon_2^{(k)}\right) = 0.$$

Therefore using the fact that the derivatives of the characteristic function at the origin under standard regularity conditions are equal to the moments of random variable, we can write

$$\begin{aligned} \frac{\partial}{\partial u_1}\psi_k\left(0, u_2\right) =& E\left[\left(ix^{(*k)} + i\varepsilon_1^{(k)}\right) e^{iu_2\left(x^{(*k)}+\varepsilon_2^{(k)}\right)}\right] \\ =& E\left[\left(ix^{(*k)}\right) e^{iu_2 x^{(*k)}}\right] Ee^{iu_2\varepsilon_2^{(k)}}. \end{aligned} \tag{17}$$

In the last equality we also make use of the independence between $x^{(*k)}$ and $\varepsilon_2^{(k)}$. These expressions also utilize the assumption of statistical independence of the measurement errors in the two proxy variables. Next note also that

$$\psi_k\left(0, u_2\right) = Ee^{iu_2 x^{(*k)}} Ee^{iu_2\varepsilon_2^{(k)}}.$$

Therefore we can write

$$\frac{\frac{\partial}{\partial u_1}\psi_k\left(0,u_2\right)}{\psi_k\left(0,u_2\right)}=\frac{E\left(ix^{(*k)}\right)e^{iu_2x^{(*k)}}}{Ee^{iu_2x^{(*k)}}}.$$

But the right hand side of the above formula is also

$$\frac{d}{du_2}\log\phi_x^{(*k)}\left(u_2\right)=\frac{d}{du_2}\log Ee^{iu_2x^{(*k)}}.$$

Since $\log\phi_x^{(*k)}\left(0\right)=0$, we can write

$$\log\phi_x^{(*k)}\left(t\right)=\int_0^t\frac{d}{du_2}\log\phi_x^{(*k)}\left(u_2\right)du_2=\int_0^t\frac{\frac{\partial}{\partial u_1}\psi_k\left(0,u_2\right)}{\psi_k\left(0,u_2\right)}du_2,$$

which immediately implies the first relation in (16):

$$\phi_x^{(*k)}\left(t\right)=\exp\left[\int_0^t\frac{\frac{\partial}{\partial u_1}\psi_k\left(0,u_2\right)}{\psi_k\left(0,u_2\right)}du_2\right].\tag{18}$$

The other two relations in (16) follow immediately from the fact that

$$\psi_{z_1}^{(k)}\left(t\right)=\psi_k\left(t,0\right)\quad\text{and}\quad\psi_{z_2}^{(k)}\left(t\right)=\psi_k\left(0,t\right)$$

and the assumption about the independence of measurement errors in two proxy variables.

To briefly summarize the results so far we should note that the expressions in (16) represent the characteristic functions of the latent vector of explanatory variables $\mathbf{x}^*$ and the observation errors ε in terms of the joint characteristic function of the observable proxy variables. In this way we can completely describe the marginal distributions of $\mathbf{x}^*$ and ε and, by independence assumption, obtain a complete description of the joint distribution of the unobservable variables. This describes the main idea of Li (2002).

Given the estimates of the characteristic functions of the latent regressor and the measurement errors, we can obtain the conditional distribution of the latent regressor given the observable proxy variables. This distribution conditional distribution for the random vector $\mathbf{x}^*$ given the vectors of observable proxy variables $f_{\mathbf{x}^*|\mathbf{z}_l}(\mathbf{x}^*|\mathbf{z}_l)$, $l=1,2$ can be written as:

$$f_{\mathbf{x}^*|\mathbf{z}_l}(\mathbf{x}^*|\mathbf{z}_l)=\frac{f_{\mathbf{x}^*}(\mathbf{x}^*)\Pi_{k=1}^K f_{\varepsilon_l}^{(k)}(z_l^{(k)}-x^{(*k)})}{f_{\mathbf{z}_l}(\mathbf{z}_l)}.$$

Then we cab obtain the marginal densities of the observable proxy variables $f_{\mathbf{z}_l}\left(\mathbf{z}_l\right)$ by the inverse Fourier transform of the joint characteristic function of the components of the vector of proxies $\mathbf{z}_l$:

$$f_{\mathbf{z}_l}\left(\mathbf{z}_l\right) = \left(\frac{1}{2\pi}\right)^k \int\limits_{-\infty}^{+\infty} \psi_{\mathbf{z}_l}(\mathbf{t}) e^{-\mathbf{z}_l'\mathbf{t}} d\mathbf{t}.$$

Next $f_{\mathbf{x}^*}(\mathbf{x}^*)$ can be determined from applying the inverse Fourier transformation to the joint characteristic function of the components of the latent explanatory variable $\mathbf{x}^*$:

$$\phi_{\mathbf{x}^*}\left(t_1, \cdots, t_K\right) = \frac{\psi_{\mathbf{z}_l}\left(t_1, \cdots, t_K\right)}{\prod\limits_{k=1}^{K} \phi_{\varepsilon_l^{(k)}}(t_k)}.$$

Let us now analyze the possibility of empirical implementation of the suggested methodology. Given n independent observations of $(\mathbf{z}_1, \mathbf{z}_2)$, the joint characteristic function of the sample $\psi_{\mathbf{z}_l}(\cdot)$ is equal to the product of characteristic functions of individual observations; and it can be estimated using its empirical analog

$$\hat{\phi_{\mathbf{z}_l}}(t_1, \cdots, t_K) = \frac{1}{2n} \sum_{l=1}^{2} \sum_{j=1}^{n} \exp\left(\sum_{k=1}^{K} i t_k z_{lj}^{(k)}\right).$$

A significant problem with this empirical characteristic function is that the inverse Fourier transformation cannot be correctly defined unless we "trim" its support. In fact, the complex exponential in the empirical characteristic function will be offset by the complex exponential in the inverse Fourier transform which will make the integral with the infinite bounds diverge. The "truncated" version of the Fourier transformation, however, will be well defined as long as the truncation parameter is finite. As a result the expression for the truncated inverse Fourier transformation to obtain the marginal density of the sample of observable proxy variables $\hat{f}_{\mathbf{z}_l}(\cdot)$ is:

$$\hat{f}_{\mathbf{z}_l}\left(z_l^{(1)}, \cdots, z_l^{(K)}\right) = \left(\frac{1}{2\pi}\right)^K \int\limits_{-T_n}^{T_n} \cdots \int\limits_{-T_n}^{T_n} e^{-i\sum_{k=1}^{K} t_k z_l^{(k)}} \hat{\phi}_{\mathbf{z}}\left(t_1, \cdots, t_K\right) dt_1, \cdots dt_K,$$

where T_n is a "trimming" parameter which is closely related to the bandwidth parameter in kernel smoothing methods (see Li (2002) for details).

To estimate the marginal density of the measurement error we need to use the formula (16) for its characteristic function from the characteristic function of the k-th component of the proxy variable and the characteristic function of the k-th component of latent vector $\mathbf{x}^*$. Namely evaluating the characteristic function for the k-th component of $\mathbf{z}$ as:

$$\widehat{\psi}_k\left(u_1,\, u_2\right) = \frac{1}{n}\sum_{j=1}^{n} \exp\left(iu_1 z_{1j}^{(k)} + iu_2 z_{2j}^{(k)}\right),$$

we can obtain the characteristic function for the k-th component of $\mathbf{x}^*$ as

$$\widehat{\phi}_{x^{(*k)}}\left(t\right) = \exp \int\limits_0^t \frac{\partial \widehat{\psi}\left(0, u_2\right)/\partial u_1}{\widehat{\psi}\left(0, u_2\right)}\, du_2,$$

and the characteristic function for the measurement error can be expressed as:

$$\widehat{\phi}_{\varepsilon^{(k)}}(t) = \frac{\widehat{\psi}_k(t,\, 0)}{\widehat{\phi}_{x^{(*k)}}(t)}.$$

Then we can obtain the density $\hat{f}_{\varepsilon^{(k)}}(\varepsilon^{(k)})$ from the truncated version of the inverse Fourier transform suggested above. Finally, using the expression for the joint characteristic function of the latent variable $\mathbf{x}^*$ in terms of the characteristic function of the proxy variables and the characteristic functions of the measurement errors, we can obtain the estimate of the density of the unobservable regressors $\hat{f}_{\mathbf{x}^*}(\cdot)$ from the corresponding empirical characteristic function. We note that the pointwise convergence of the estimated density to the true density of the latent regressor is established under additional assumptions, which restrict the densities to have finite supports and require that the characteristic functions are uniformly bounded by exponential functions and integrable on the support.

Given the first step nonparametric estimator $\hat{f}_{\mathbf{x}^*|\mathbf{z}_l}(\mathbf{x}^*|\mathbf{z})$, a semiparametric nonlinear least-squares estimator $\hat{\beta}$ for β can be obtained by minimizing:

$$S_{SP} = \frac{1}{n}\sum_{l=1}^{2}\sum_{i=1}^{n}[y_i - \int g(\mathbf{x}^*;\beta)\hat{f}_{\mathbf{x}^*|\mathbf{z}_l}(\mathbf{x}^*|\mathbf{z}_{li})d\mathbf{x}^*]^2.$$

Li (2002) establishes the uniform convergence (with rate) of the nonparametric estimator $\hat{f}_{\mathbf{x}^*|\mathbf{z}_l}(\mathbf{x}^*|\mathbf{z})$ to the true conditional density $f_{\mathbf{x}^*|\mathbf{z}_l}(\mathbf{x}^*|\mathbf{z}_l)$, as well as the consistency of $\hat{\beta}$ to the true parameters of interest β. The method of Li (2002) can be readily extended to any nonlinear EIV models as long as there are repeated measurement available in the sample; see e.g. Li and Hsiao (2004) for consistent estimation of likelihood-based nonlinear EIV models.

Recently **Schennach (2004a)** introduces a somewhat different solution to the problem of recovering the density of latent variable in nonlinear model with classical measurement errors. Schennach (2004a) considers the following model (we follow the previous notations for the sake of continuity):

$$y = \sum_{k=1}^{M} \beta_k h_k(x^*) + \sum_{j=1}^{J} \beta_{j+M} \omega_l + u,$$

where y and ω_j, for $j = 1, \cdots, J$, can be observed, while x^* is the unobserved latent variable with two observable measurements z_1 and z_2:

$$z_l = x^* + \epsilon_l, \quad l = 1, 2,$$

and the measurement errors are ϵ_1 and ϵ_2, u is the disturbance. For convenience, set $\omega_0 = y$ and use ω_j, $j = 0, \cdots, J$, to represent all the observed variables.

Schennach (2004a) relaxes the strong independence assumptions between the measurement errors and only mean independence is required:

$$E[u \mid x^*, \epsilon_2] = 0,$$

$$E[\epsilon_1 \mid x^*, \epsilon_2] = 0, \tag{19}$$

$$E[\omega_j \mid x^*, \epsilon_2] = E[\omega_j \mid x^*], \quad \text{for } j = 1, \cdots, J. \tag{20}$$

However, the independence between ϵ_2 and x^* is reserved, indicating that we are still considering a classical measurement error problem.

The estimation procedure can be divided into two parts: the least square estimation part where the parameters for the observable variable are obtained and the part dealing with measurement errors. Given the specified model, the objective function of least square minimization is:

$$E\left[y - \sum_{k=1}^{M} \beta_k h_k(x^*) + \sum_{j=1}^{J} \beta_{j+M} \omega_j\right]^2.$$

Clearly, the vector of coefficients β can be identified if the second moments $E[\omega_j \omega_{j'}]$, for j and $j' = 0, 1, \cdots, J$, $E[h_k(x^*)h_{k'}(x^*)]$, for k and $k' = 1, \cdots, M$, and $E[\omega_j h_k(x^*)]$ for $j = 0, 1, \cdots, J$, and $k = 1, \cdots, M$ are known. Since ω_j is observable, its second moment $E[\omega_j \omega_{j'}]$ can be estimated by its sample counterpart. However, the two moments $E[h_k(x^*)h_{k'}(x^*)]$ and $E[\omega_j h_k(x^*)]$ depend on the unobservable latent variable x^* which is not directly observed in the sample without measurement error. Schennach (2004a) demonstrates that, by making use of the characteristic function approach, the distribution of x^* and therefore these moments can be related to the sample distribution of the two observable measurements of x^*. The key point here is again to derive the characteristic function of x^* and the joint features of this characteristic function with other observable variables from sample information.

All the moments required above have the form of $E\left[W\gamma\left(x^*\right)\right]$ where $W = 1$ when $\gamma\left(x^*\right)$ is one of $h_k(x^*)h_{k'}(x^*)$, and $W = w_j, j = 0, \ldots, J$ when $\gamma\left(x^*\right)$ is one of $h_k\left(x^*\right)$. Theorem 1 in Schennach (2004a) shows that this moment $E\left[W\gamma\left(x^*\right)\right]$ can be recovered from observable sampling information through

$$E[W\gamma(x^*)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mu_\gamma(-\chi)\phi_W(\chi)d\chi, \tag{21}$$

where

$$\phi_W(\chi) \equiv E\left[We^{i\chi x^*}\right] = \frac{E[We^{i\chi z_2}]}{E[e^{i\chi z_2}]}\exp\left(\int_0^\chi i\frac{E[z_1 e^{i\zeta z_2}]}{E[e^{i\zeta z_2}]}d\zeta\right), \tag{22}$$

and $\mu_\gamma(-\chi)$ is the Fourier transformation of $\gamma(x^*)$ defined as

$$\mu_\gamma(-\chi) = \int e^{-i\chi x^*}\gamma\left(x^*\right)dx^*.$$

To understand this theorem we need to first understand the relation in (21) where $\phi_W(\chi) \equiv E\left[We^{i\chi x^*}\right]$. Next we will see how $\phi_W(\chi)$ can be written as the last term in (22).

For further manipulations with (21), first recall the definition the Dirac's δ - function $\delta\left(x^*\right)$. δ - function is formally defined as a functional in the space of test functions $f \in \mathcal{D}$ which are infinitely differentiable with finite support of all derivatives. Then by definition for all $f \in \mathcal{D}$ the function $\delta\left(x^* - a\right)$ is a continuous linear functional on $\mathcal{D}$ such that:

$$\delta_a * f = \int\limits_{-\infty}^{+\infty} \delta\left(x^* - a\right) f(x^*)\, dx^* = f(a).$$

The continuous linear functionals mapping from $\mathcal{D}$ to $\mathbb{R}$ are usually called the generalized functions. In this way the result of the Fourier transformation is similar to the application of the δ-function and in a shorthand notation we can write:

$$\int e^{ix^*\chi} d\chi = \delta\left(\frac{x^*}{-2\pi}\right) = 2\pi\delta\left(x^*\right),$$

implying that the result of the application of the linear functional corresponding to the δ-function is the same as the application of the corresponding Fourier transformation. This transformation might only exist as a generalized function instead of a regular function.

Then to show (21), we begin with its right hand side using the definition of $\mu_\gamma(-\chi)$ and $\phi_W(\chi)$:

$$\begin{aligned}
&\frac{1}{2\pi}\int \mu_\gamma(-\chi)\phi_W(\chi)d\chi = \frac{1}{2\pi}\int\left[\int e^{-i\chi x^*}\gamma\left(x^*\right)dx^*\right]\left[\int\int We^{i\chi\tilde{x}^*}f\left(W,\tilde{x}^*\right)d\tilde{x}^*dW\right]d\chi \\
=&\frac{1}{2\pi}\int\int W\gamma\left(x^*\right)\int\int e^{i\chi\left(\tilde{x}^*-x^*\right)}d\chi f\left(W,\tilde{x}^*\right)d\tilde{x}^*dWdx^* \\
=&\int\int W\gamma\left(x^*\right)\int\delta\left(\tilde{x}^*-x^*\right)f\left(W,\tilde{x}^*\right)d\tilde{x}^*dWdx^* \\
=&\int\int W\gamma\left(x^*\right)f\left(W,x^*\right)dWdx^* = E\left[Wu\left(x^*\right)\right].
\end{aligned}$$

Next we consider showing the second equality in (22). Note first that we can write

$$E\left[We^{i\chi x^*}\right] = \frac{E\left[We^{i\chi x^*}\right]}{E\left[e^{i\chi x^*}\right]}E\left[e^{i\chi x^*}\right].$$

The last term $E\left[e^{i\chi x^*}\right]$ follows from the same derivations from (16) to (18), where it is noted that assumption (19) is sufficient for equation (17) to hold. Then (18) can be restated as

$$E\left[e^{i\chi x^*}\right] = \phi_{x^*}\left(\chi\right) = \exp\left(\int_0^\chi i\frac{E[z_1e^{i\zeta z_2}]}{E[e^{i\zeta z_2}]}d\zeta\right).$$

Finally to show $\frac{E\left[We^{i\chi x^*}\right]}{E\left[e^{i\chi x^*}\right]} = \frac{E\left[We^{i\chi z_2}\right]}{E\left[e^{i\chi z_2}\right]}$, consider the right hand side

$$\frac{E\left[We^{i\chi z_2}\right]}{E\left[e^{i\chi z_2}\right]} = \frac{E\left[We^{i\chi(x^*+\epsilon_2)}\right]}{E\left[e^{i\chi(x^*+\epsilon_2)}\right]} = \frac{E\left[We^{i\chi x^*}\right]E\left[e^{i\chi\epsilon_2}\right]}{E\left[e^{i\chi x^*}\right]E\left[e^{i\chi\epsilon_2}\right]} = \frac{E\left[We^{i\chi x^*}\right]}{E\left[e^{i\chi x^*}\right]}.$$

The second equality above follows from assumption (20) and the independence between x^* and ϵ_2. This completes the proof for (21) and (22). Note that when $W \equiv 1$, the first term in (22) vanishes and $\phi_W(\chi) = \phi_{x^*}(\chi)$ is just the characteristic function for x^*.

To summarize, the equations (21) and (22) provide us with the tool to recover the characteristic function for the latent variable x^* from the characteristic functions of observable proxy variables. Given sampling information about y, w_j, z_1, z_2, one can form sample analogs of the population expectations in (21) and (22), and use them to form the estimates for $E[\omega_j \omega_{j'}]$, $E[h_k(x^*)h_{k'}(x^*)]$ and $E[\omega_j h_k(x^*)]$, which are then used to compute the least square estimator. Asymptotic theory for this estimator is developed in Schennach (2004a).

The estimation procedure described above is a generalization of previous research in polynomial and linear models. If $h_k(x^*)$ is a polynomial, as the case considered in Hausman, Newey, Ichimura, and Powell (1991), given the standard assumptions about the distributions under considerations, the moments of interest $E[W\gamma(x^*)]$ reduce to:

$$E[W\gamma(x^*)] = (-i)^s \frac{d^s \phi_W(\chi)}{d\chi^s}\bigg|_{\chi=0}. \tag{23}$$

which can be used to derive the same estimates as in Hausman, Newey, Ichimura, and Powell (1991). Furthermore, in case of a linear model, this approach is equivalent to the linear IV estimation method.

The estimation approach for multivariate measurement errors is also considered in Schennach (2004a). It is analogous to the univariate case, however, it extends to a more general class of M-estimators. Let the unobservable variable $\mathbf{x}^*$ be a $K \times 1$ random vector and $\mathbf{z}_1$ and $\mathbf{z}_2$ be the corresponding repeated measurements (proxy variables) for $\mathbf{x}^*$, such that $\mathbf{z}_l = \mathbf{x}^* + \varepsilon^l$, $l = 1, 2$. To disentangle the characteristic function of the latent vector $\mathbf{x}^*$ we still need to assume the mean independence between $\mathbf{x}^*$ and $\mathbf{z}_1$ $\mathbf{z}_2$: let $z_l^{(k)}$ represent the k-th element of proxy variable $\mathbf{z}_l$, then for each k, such that $k' \in (1, \cdots, K)$, $k \neq k'$, assume the mean independence for the components of the vector of measurement errors for the first proxy variable $E[\varepsilon_1^{(k)} \mid x^{(*k)}, \varepsilon_2^{(k)}] = 0$ and complete statistical independence for components in the vector of measurement errors in the second proxy vector $\epsilon_2^{(k)}$ from the latent vector of explanatory variables $\mathbf{x}^*$, the observable vector of explanatory variables $\mathbf{w}$ and $\epsilon_2^{(k')}$ for the other components $k' = 1, \ldots, K$ and $k' \neq k$. The nonlinear model can take a general form as an M-estimator with the kernel $R(\mathbf{x}^*, \mathbf{w}, \beta)$. The problem for M-estimator can be defined as:

$$\widehat{\beta} = \underset{\beta \in \mathcal{B}}{\operatorname{argmax}}\, H\left(\widetilde{E}\left[R\left(\mathbf{x}^*,\ \mathbf{w},\ \beta\right)\right]\right),$$

where $H(\cdot)$ is a general non-linear function and $\widetilde{E}\left[\cdot\right]$ is the estimate of the corresponding expectation.

The construction of such M-estimator requires information on each moment $E[R_j(\mathbf{x}^*, \mathbf{w}, \beta)]$, $j = 1 \cdots J$ and J is the dimension of function $R(\cdot)$. To unify the notation, denote each $R_j(\mathbf{x}^*, \mathbf{w}, \beta)$ by $\gamma(\mathbf{x}^*, \mathbf{w}, \beta)$. The evaluation of the estimate of the expectation $E[\gamma(\mathbf{x}^*, \mathbf{w}, \beta)]$ proceeds in two steps:

Step 1: express $\gamma(\mathbf{x}^*, \mathbf{w}, \beta)$ as a linear combination of chosen basis functions and determine the set of weights $\mu(\chi, \omega, \beta)$ for the expansion of the function $\gamma(\cdot)$ in the chosen basis.

Schennach (2004a) uses the separable basis functions for the expansion of $\gamma(\cdot)$. The separable part for the vector of latent regressors $\mathbf{x}^*$ is spanned by the Fourier basis in the form $e^{-i\chi\mathbf{x}^*}$. The separable part for the vector of observable regressors $\mathbf{w}$ remains unspecified for the non-parametric flexibility and is represented by a general function $b_\omega(\mathbf{w})$. The indices for the basis functions $b_\omega(\cdot)$ belong to some finite-dimensional index set $\mathcal{W}$. It is important that the basis functions for $\mathbf{x}^*$ and $\mathbf{w}$ are separable. This assumption will allow us to make the necessary manipulations with the characteristic functions in the next step. Assuming the completeness of the chosen basis functions in the space containing the parametric family $\gamma(\cdot, \beta)$, given components of expansion $\mu(\chi, \omega, \beta)$ in the chosen separable basis, $\gamma(\mathbf{x}^*, \mathbf{w}, \beta)$ can be expressed as:

$$\gamma\left(\mathbf{x}^*, \mathbf{w}, \beta\right) = \left(\frac{1}{2\pi}\right)^{-K} \sum_{\omega \in \mathcal{W}} \int \cdots \int \mu(-\chi, \omega, \beta) e^{i\chi\mathbf{x}^*} b_\omega(\mathbf{w}) d\chi_1 \cdots d\chi_K. \tag{24}$$

If the chosen basis function for $\mathbf{w}$ is also continuous, we can further replace the summation by integral for $\mathbf{w}$. Finally, the weights $\mu(\chi, \omega, \beta)$ can be solved out by rearranging equation (22) and working out all the summation and integrals.

Step 2: using the first step result $\mu(\chi, \omega, \beta)$, derive $E[\gamma(\mathbf{x}^*, \mathbf{w}, \beta)]$ based on the observations of the proxy vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ for $\mathbf{x}^*$.

In order to do the estimation we need to make additional assumptions, other than the mean independence assumption for $\mathbf{x}^*$ and $\mathbf{z}_1$ $\mathbf{z}_2$. Specifically we require that $E\left[|x^{(*k)}|\right]$, $E\left[|\epsilon_1^{(k)}|\right]$, for $k = 1, \cdots, K$ are finite. Moreover for all indices ω in $\mathcal{W}$ we require that $E\left[|b_\omega(\mathbf{w})|\right]$ is bounded. In these circumstances if the expectation $E\left[\gamma\left(\mathbf{x}^*, \mathbf{w}, \beta\right)\right]$ exists, then it can be expressed as:

$$E\left[\gamma\left(\mathbf{x}^*, \mathbf{w}, \beta\right)\right] = \left(\frac{1}{2\pi}\right)^{-K} \sum_{\omega \in \mathcal{W}} \int \cdots \int \mu(-\chi, \omega, \beta)\phi_b(\chi, \omega) d\chi_1 \cdots d\chi_K, \tag{25}$$

where

$$\begin{aligned} &\phi_b(\chi, \omega) = e^{i\chi\mathbf{x}^*} b_\omega(\mathbf{w}) = \\ &= E\left[b_\omega(w) e^{i\chi\mathbf{z}_2}\right] \left(\prod_{k=1}^{K} E\left[e^{i\chi_k z_2^{(k)}}\right]\right)^{-1} \prod_{k=1}^{K} \exp\left(\int_0^{\chi_k} \frac{iE\left[z_1^{(k)} e^{i\zeta_k z_2^{(k)}}\right]}{E\left[e^{i\zeta_k z_2^{(k)}}\right]} d\zeta_k\right). \end{aligned} \tag{26}$$

Note that the second component of the kernel in the integral (25), $\mu(\chi, \omega, \beta)$ has been determined in the first step from the components of the expansion if $\gamma(\cdot)$ in the chosen separable basis. The second component represents the element of the basis corresponding to the coordinate $\mu(\cdot)$. In the this function is represented using the approach which we applied in the one-dimensional case. See Schennach (2004a) section 3.3 for details about the asymptotic properties of the estimator.

Schennach (2004a) applies the deconvolution technique to analyze Engel curves of households using data from the Consumer Expenditure Survey. The Engel curve describes the dependence of the proportion of income spent on a certain categories of goods on the total expenditure. The author assumes that the total expenditure is reported with error. To reduce the bias in the estimates due to the observational error, the author uses two alternative estimates of the total expenditure. The first estimate is the expenditure reported for the household in the current quarter, while the second estimate is the expenditure reported in the next quarter. The author compares the estimates obtained using the characteristic function approach and the standard feasible GLS estimates. Her estimates show that the FGLS - estimated elasticities of expenditure on groups of goods with respect to the total expenditure are lower than the elasticities obtained using the deconvolution technique that she provided. This can suggest that the method of the author corrects the downward bias in the estimates of income elasticity of consumption that arises from the errors in the observed total expenditure.

The deconvolution method via repeated measurement can in fact allow for fully nonparametric identification and estimation of models with classical measurement errors with unknown error distributions. See e.g., Li and Vuong (1998), Li, Perrigne, and Vuong (2000), Schennach (2004b) and Bonhomme and Robin (2006).

3.4 Nonlinear EIV models with strong instrumental variables

Although the standard IV assumption (for a linear EIV model) is not enough to allow for point identification of the parameters in a general nonlinear EIV model, some slightly stronger notions of IVs do imply point identification and consistent estimation. In fact, the methods using double measurements discussed in the last subsection could be regarded as special forms of IVs. In this subsection we shall review some additional IV approaches.

3.4.1 Nonlinear EIV models with generalized double measurements

Carroll, Ruppert, Crainiceanu, Tosteson, and Karagas (2004) consider a general nonlinear regression model with a mismeasured regressor and a valid instrument, in which the regression form could even be fully nonparametric. In this paper the dependent variable y is a function of the latent true regressor x^* and a vector of observed covariates v. x^* is mismeasured as x, and there is also an instrument z available for the mismeasured regressor x. z follows a varying-coefficient model that is linear in x^* with coefficients being smooth functions of v; hence in some sense z could be regarded as a generalized notion of second measurement of latent variable x^*.

Without covariates, the simplest specification considered by the authors is given by:

$$\begin{aligned}
&y = m\left(x^*\right) + \epsilon,\ E(\epsilon) = 0,\\
&x = x^* + \eta,\ E(\eta) = 0,\\
&z = \alpha_0 + \alpha_1 x^* + \zeta,\ E(\zeta) = 0,\ \alpha_1 \neq 0.
\end{aligned}$$

Under the assumptions that $(x^*,\, \eta,\, \epsilon,\, \zeta)$ are *mutually uncorrelated*, and that

$$cov\left\{x^*,\, m\left(x^*\right)\right\} \neq 0,$$

the authors prove that the parameters α_0, α_1, $E(x^*)$, $Var(x^*)$, $Var(\eta)$ and $Var(\zeta)$, as well as the unknown conditional mean function $m(x^*)$ are all identified.

For some classes of functions $m(x^*)$, the assumption $cov\{x^*,\, m(x^*)\} \neq 0$ might fail. The authors then point out this assumption can be weakened to: there exists some positive integer k such that

$$cov\left\{[x^* - E(x^*)]^k,\, m(x^*)\right\} \neq 0,$$

but the mutual uncorrelatedness of $(x^*,\, \eta,\, \epsilon,\, \zeta)$ assumption has to be strengthened to *mutual independence.*

More specifically, they assume that for a fixed K there are $2K$ finite moments of the vector of observable variables $(y,\, x,\, z)$ and for some (unknown) natural $k \leq K$:

$$\rho_k = cov\left\{m(x^*),\, [x^* - E(x^*)]^k\right\} = cov\left\{y,\, [x - E(x)]^k\right\} \neq 0.$$

Once the number k is obtained, the slope coefficient in the "instrument" equation is identified as:

$$\alpha_1 = sign\{cov(x,\, z)\}\, |\frac{cov\left[y,\, [z - E(z)]^k\right]}{\rho_k}|^{1/k}.$$

The estimation procedure suggested by the authors is based on testing zero correlation ρ_k and then using it to form expressions for the slope in the "instrument" equation. The estimate of k is determined as the first number for which the hypothesis of equality to zero is rejected, or (if the null is never rejected) this is the number corresponding to the smallest p-value.

The more general model considered by the authors contains the following equations:

$$\begin{aligned} y &= g(v, x^*, \epsilon) \\ x &= x^* + \eta \\ z &= \alpha_0(v) + \alpha_1(v)x^* + \zeta \end{aligned}$$

The set of observable variables includes y, x, z and v, where v is the set of covariates observed without error. Such model includes several classes of models such as generalized linear regression model.

The identification assumption of the general model is that the error terms ϵ, η, and ζ are mutually independent and are independent from the covariates v and x^*. An additional assumption is that the error-free covariate v is univariate with support on [0, 1] and its density is bounded away from zero on the support. In addition, they assume that there is a known bound $L \geq 1$ such that for some positive integer $1 \leq l \leq L$, the conditional covariance:

$$cov \left\{ y,\, (x - E(x \mid v))^k \mid v \right\},$$

is zero for all $k < l$ and is bounded away from zero for all v in the support if $k = l$. Finally, the authors assume that the slope coefficient in the "instrument" equation α_1 is constant.

Under these assumptions one can recover the slope coefficient in the "instrument" equation from the ratio of the covariances:

$$\alpha_1^l = \frac{cov\left\{ y,\, (z - E(z \mid v))^l \mid v \right\}}{cov\left\{ y,\, (x - E(x \mid v))^l \mid v \right\}}.$$

The slope coefficient can be estimated by first non-parametrically estimating the covariances of interest. Then, choosing the appropriate trimming points on the support of v, we can obtain the estimate of the slope coefficient as a trimmed average over the observations.

Once the parameters of the "instrument" equation are estimated, they can be used to recover the true regression function by running a non-parametric regression. This requires a mild technical assumption that the kernel estimate of the conditional expectation $m_{jkp}(v) = E\left(y^j x^k z^p \mid v\right)$ can approximate the true conditional expectation $m_{jkp}(v)$ uniformly well over $v \in [a, b]$ for $0 < a < b < 1$. More precisely, for a kernel function $K_h(\cdot)$, they assume that the approximation error can be written:

$$\widehat{m}_{jkp}(v) - m_{jkp}(v) = \left[\frac{1}{n f_v(v)} \sum_{i=1}^{n} K_h(v_i - v)\, u_{jkpi} \right] + O_p\left\{ n^{-2/3} \log n \right\}.$$

Here $f_v(\cdot)$ is the density of v, and $E(u_{jkpi} \mid v_i) = 0$, while $var(u_{jkpi} \mid v_i) \leq A < \infty$.

Under these assumptions the authors prove that the parameters in the "instrument" equation and the estimates of the variances of errors will be $\sqrt{n}$ consistent. For practical purposes the authors recommend use trimming of the support of v for estimation.

The authors then extend the analysis to the case when the coefficient α_1 depends on the covariate v. It can be recovered from the non-parametric estimates for the covariances by taking their ratio. Under these assumptions the authors prove that the main regression function can be non-parametrically estimated from the observed variables (y, x, z, v). As additional methods for estimation, the authors suggest using deconvolution kernels, penalized splines, SIMEX method, or Bayesian penalized splines estimator.

The authors illustrate their estimation procedure using examples from two medical studies. The first study focuses on the analysis of the effect of arsenic exposure on the development of skin, bladder, and lung cancer. The measurement error comes from the fact that physical arsenic exposure (through water) does not necessarily imply that the exposure is biologically active. The application of the suggested method allows the authors to find the effect of the biologically active arsenic exposure on the frequency of cancer incidents. In the other example the authors study the dependence between cancer incidents and diet. The measurement error comes from the fact that the data on the protein and energy intake are coming from the self-reported food frequency questionnaires, which can record the true food intake with an error. The estimation method suggested in the paper can allow the authors to estimate the effect of the structure of the diet on the frequency of related cancer incidents.

3.4.2 Nonlinear EIV models of generalized Berkson type

In statistics, medical and biology literature, there is a special class of measurement error models, called **Berkson models**, in which the latent true variable of interest x^* is predicted (or caused) by the observed random variable z via the causal equation:

$$x^* = z + \zeta,$$

where the unobserved random measurement error ζ is assumed to be independent of the observed predictor z. See e.g., Fuller (1987) and Carroll, Ruppert, and Stefanski (1995) for motivations and explanations of the Berkson-error models; and Wang (2004) for a recent identification and estimation of nonlinear regression model with Berkson measurement errors.

Although the Berkson-error model might not be a realistic measurement error model to describe many economic data sets, the idea that some observed random variables predict latent true variable of interest might still be sensible in some economics applications.

Newey (2001) considers the following form of a nonlinear EIV regression model with classical error and a causal (prediction) equation:

$$\begin{aligned} y &= f\left(x^{*}, \delta_0\right) + \epsilon, \\ x &= x^{*} + \eta, \\ x^{*} &= \pi_0' z + \sigma_0 \zeta, \end{aligned}$$

where the errors are conditionally mean independent: $E\left[\epsilon \mid z, \zeta\right] = 0$ and $E\left[\eta \mid z, \epsilon, \zeta\right] = 0$. The measurement equation $x = x^{*} + \eta$ contains the classical measurement error η (i.e., x^{*} and η are statistically independent. The unobserved prediction error ζ and the "predictor" z in the causal equation $x^{*} = \pi_0' z + \sigma_0 \zeta$ are assumed to be statistically independent. The vector z is assumed to contain a constant; hence the prediction error ζ is normalized to have zero mean and identity covariance matrix. Apart from the restrictions on the means and variances, no parametric restrictions are imposed on the distributions of the errors. The parameters of interest are $(\delta_0, \pi_0, \sigma_0)$. This model has also been studied in Wang and Hsiao (1995), who proposed similar identification assumptions but a different estimation procedure.

The model assumptions allow one to write the moment equations for conditional expectations of y given z, the product $y\,x$ given z and the regressor x given z in terms of the unknown density of the prediction error ζ. If we denote this density by $g_0(\zeta)$, then we obtain the following three sets of conditional moment restrictions:

$$E\left[y \mid z\right] = E\left[f\left(x^{*}, \delta_0\right) | z\right] = \int f\left(\pi_0' z + \sigma_0 \zeta,\, \delta_0\right) g_0\left(\zeta\right)\, d\zeta, \tag{27}$$

$$E\left[y\,x \mid z\right] = \int \left[\pi_0' z + \sigma_0 \zeta\right] f\left(\pi_0' z + \sigma_0 \zeta,\, \delta_0\right) g_0\left(\zeta\right)\, d\zeta, \tag{28}$$

and

$$E\left[x \mid z\right] = \pi_0' z. \tag{29}$$

Newey (2001) suggests a Simulated Method of Moments (SMM) to estimate the parameters of interest $(\delta_0, \pi_0, \sigma_0)$ and the nuisance function g_0 (the density of the prediction error ζ). To do so, assume that we can simulate from some density $\varphi(\zeta)$. Then represent the density of the error term as:

$$g(\zeta, \gamma) = P(\zeta, \gamma)\, \varphi(\zeta),$$

where $P(\zeta, \gamma) = \sum_{j=1}^{J} \gamma_j p_j(\zeta)$ for some basis functions $p_j(\cdot)$. The coefficients in the expansion should be chosen so that $g(\zeta, \gamma)$ is a valid density. The coefficient choices need to be normalized to impose restrictions on the first two moments of this density. One possible way of imposing such restrictions is to add them as extra moments into the original system of moments.

In the next step, Newey (2001) construct a system of simulated moments $\widehat{\rho}(\alpha)$ for $\widehat{\rho}(\alpha)$ for $\alpha = (\delta', \sigma, \gamma')'$ as:

$$\widehat{\rho}_i(\alpha) = \begin{pmatrix} y_i \\ L x_i y_i \end{pmatrix} - \frac{1}{S} \sum_{s=1}^{S} \begin{pmatrix} f(\pi' z_i + \sigma \zeta_{is}, \delta) \\ L(\pi' z_i + \sigma \zeta_{is})\, f(\pi' z_i + \sigma \zeta_{is}, \delta) \end{pmatrix} P(\zeta_{is}, \gamma)$$

where L is the matrix selecting the regressors containing the measurement error.

This system of moments can be used to form a method of moments objective. Specifically, if $\widehat{A}(z_i)$ is a vector of instruments for the observation i then the sample moment equations will take the form:

$$m_i(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \widehat{A}(z_i) \widehat{\rho}_i(\alpha).$$

The weighting matrix can be obtained from a preliminary estimate for the unknown parameter vector. The standard GMM procedure then follows. Newey (2001) shows that such a procedure will produce consistent estimates of the parameter vector under a set of regularity conditions. Notice that the system of three conditional moment equations (27, 28, 29) and the estimation procedure fit into the framework studied in Ai and Chen (2003), whose results are directly applicable to derive root-n asymptotic normality and consistent asymptotic variance estimator of Newey's estimator for $(\delta_0, \pi_0, \sigma_0)$.

The suggested estimator is then applied to the estimation of Engel curves as a dependence between the share on a specific commodity group from income. The dependence is specified in the form of a dependence with the logarithm and an inverse of individual income determining the right-hand side. The author assumes that the individual income is measured with an error which comes in a multiplicative form, allowing to switch to the analysis of the logarithm of income instead of the level. In estimation the author uses the data from the 1982 Consumer Expenditure Survey, giving the shares of individual expenditure on several commodity groups. The estimation method of the paper is implemented for the assumption of a Gaussian error and for the Hermite polynomial specification for the error density, and it compared with the results of the conventional Least Squares (LS) and the Instrumental Variables (IV) estimators. The estimation results show significant downward biases in the LS and IV estimates, while the suggested SMM estimates are close for both Gaussian specification and the flexible Hermite specification for the distribution of the error term. This implies that the suggested method can be an effective tool for reduction of measurement errors in the dependent variables in non-linear models.

The model studied in Newey (2001) and Wang and Hsiao (1995) are recently extended by Schennach (2006), using Fourier deconvolution techniques, to a nonparametric regression setup: $y = g\left(x^*\right) + \epsilon$, where the functional form $g(x^*)$ is unknown. The complete model can be written as:

$$\begin{aligned} y &= g\left(x^*\right) + \epsilon, \\ x &= x^* + \eta, \\ x^* &= m(w) + \zeta, \ E\left[\zeta\right] = 0, \end{aligned}$$

The imposed assumptions include mean independence $E\left[\epsilon \mid w,\, \zeta\right] = 0$, $E\left[\eta \mid w, \zeta, \epsilon\right] = 0$, and the statistical independence of ζ from w.

Given the exogeneity of the error term in the last equation, the author suggests to identify this equation by a non-parametric projection of x on v. It is then possible to substitute the last equation by $x^* = z - u$, where $z = m(w)$ and $u = -\zeta$. The system takes the form:

$$\begin{aligned} y &= g\left(x^{*}\right)+\epsilon, \\ x &= x^{*}+\eta, \\ x^{*} &= z-u. \end{aligned}$$

The new set of assumptions is the same as before with the substitution of conditioning on w with conditioning on z.

The moments in this model conditional on z can then be written in terms of the integrals over the distribution of the error term u. This leads to the system of conditional moments in the form:

$$\begin{array}{l} E\left[y \mid z\right]=\int g(z-u)\, dF(u), \\ E\left[x\, y \mid z\right]=\int (z-u)\, g(z-u)\, dF(u). \end{array}$$

The next step of the author is to write the functions under consideration in terms of their Fourier transformations. This produces the following expressions:

$$\begin{array}{l} \epsilon_{y}\left(\xi\right) \equiv \int E\left[y \mid z\right] e^{i\xi z}\, dz, \\ \epsilon_{xy}\left(\xi\right) \equiv \int E\left[x\, y \mid z\right] e^{i\xi z}\, dz, \\ \gamma\left(\xi\right) \equiv \int g\left(x^{*}\right) e^{i\xi x^{*}}\, dx^{*}, \\ \phi\left(\xi\right) \equiv \int e^{i\xi u}\, dF(u.) \end{array}$$

These expressions are related through the following system of differential equations:

$$\begin{array}{l} \epsilon_{y}\left(\xi\right)=\gamma\left(\xi\right)\phi\left(\xi\right), \\ \mathbf{i}\,\epsilon_{xy}\left(\xi\right)=\dot{\gamma}\left(\xi\right)\phi\left(\xi\right), \end{array}$$

where $\dot{\gamma}\left(\xi\right)=\frac{d\gamma}{d\xi}$. The author notes that this system might not be directly solvable. In case of discontinuities in the regression function, its Fourier transformation will contain a singular component which will invalidate "arithmetic" solutions. Only regular components of the Fourier transformation can be used for algebraic manipulations.

To solve this system of equations the author imposes additional assumptions on the distributions and the regression function. First, the moments and the regression function cannot grow faster than a polynomial rate. Second, the absolute value of the error u has a finite expectation and its characteristic function is never equal to zero. Third, the support of the characteristic function of the regression function is finite and restricted to a segment $\left[-\overline{\xi},\, \overline{\xi}\right]$. This means that $\gamma\left(\xi\right) \neq 0$ for $\xi \in \left[-\overline{\xi},\, \overline{\xi}\right]$ and $\gamma\left(\xi\right) = 0$ otherwise. The constant $\overline{\xi}$ restricting the segment can potentially be infinite.

Under these assumptions, the regression function can be determined from its Fourier transform. The Fourier transform of the regression function can be recovered from the regular components of the Fourier transforms of the moment equations (indexed by r) by the expression:

$$\gamma\left(\xi\right) = \begin{cases} 0, & \text{if } \epsilon_y\left(\xi\right) = 0 \\ \epsilon_y\left(\xi\right) \exp\left(-\int\limits_0^{\xi} \frac{\mathbf{i}\, \epsilon_{(z-x)y,\, r}(s)}{\epsilon_{y,\, r}(s)}\, ds\right), & \text{otherwise.} \end{cases}$$

This relation is derived from transforming the original system of Fourier transformations into the form of:

$$\begin{aligned} &\epsilon_y\left(\xi\right) = \gamma\left(\xi\right)\phi\left(\xi\right) \\ &\mathbf{i}\, \epsilon_{(z-x)y}\left(\xi\right) = \gamma\left(\xi\right)\dot{\phi}\left(\xi\right) \end{aligned}$$

which follows in turn from the fact that $\frac{d\epsilon_y(\xi)}{d\xi} = i\epsilon_{zy}\left(\xi\right)$. The regression function itself can be recovered from the inverse Fourier transformation of the function $\gamma\left(\xi\right)$.

The estimation method suggested by the author consists of three steps. In the first step, x is projected on w to calculate z. In the second step, the distribution of the disturbance u is estimated by kernel estimator given the projection results. In the last step, the density estimate is used to form a system of moment equations for coefficients of the Fourier transformations of the regression function and of the conditional moments of the outcome y and cross-product $y\, x$.

4 Nonlinear EIV Models With Nonclassical Errors

The recent applied economics literature has raised great concerns about the validity of the classical measurement error assumption. For example, in economic data, it is often the case that data sets rely on individual respondents to provide information. It may be hard to tell whether or not respondents are making up their answers, and more crucially, whether the measurement error is correlated with the latent true variable and some of the other observed variables. Studies by Bound and Krueger (1991), Bound, Brown, Duncan, and Rodgers (1994), Bollinger (1998) and Bound, Brown, and Mathiowetz (2001) have all documented evidences of nonclassical measurement errors in economics data sets. In this section we review some of the very recent theoretical advances on nonlinear models with nonclassical measurement errors. We first survey results on misclassification of discrete variables. We then review some current results on nonlinear models of continuous variables measured with nonclassical errors.

4.1 Misclassification of discrete variables

Measurement errors in binary or discrete variables usually take the form of *misclassification*. For example, a unionized worker might be misclassified as one who is not unionized. When the variable of interest and its measurement are both binary, the measurement error can not be independent of the true binary variable. Typically, misclassification introduces a negative correlation, or mean reversion, between the errors and the true values. As a result, using traditional estimation methods, such as probit and logit, will generate inconsistent estimates.

4.1.1 Misclassification of discrete dependent variables

To correct the misclassification in the discrete dependent variables, **Hausman, Abrevaya, and Scott-Morton (1998)** introduce a modified maximum likelihood estimator, which can consistently estimate coefficients and the explicit extent of misclassification. Suppose the binary choice model for latent variable y^* is:

$$y_i^* = x_i'\beta + \epsilon_i, \ \epsilon_i \text{ is independent of } x_i.$$

The probability distribution function of $-\epsilon_i$ is the same for all i and is denoted as F. The authors consider the binary response model where the true response is induced by zero threshold crossing of the latent variable: $\tilde{y}_i = 1(y_i^* \geq 0)$. This response is observed with misclassification, where the misclassified indicator is denoted by y_i. Let α_0 denote the probability of misclassification as one and α_1 denote the misclassification probability as zero, both of which are assumed to be independent of the covariates x_i. Then:

$$\alpha_0 = \Pr\left(y_i = 1 \mid \tilde{y}_i = 0\right) = \Pr\left(y_i = 1 \mid \tilde{y}_i = 0, x_i\right),$$
$$\alpha_1 = \Pr(y_i = 0 \mid \tilde{y}_i = 1) = \Pr(y_i = 0 \mid \tilde{y}_i = 1, x_i).$$

As a result, the expected value of the observed dependent variable given the misclassification probabilities can be specified as:

$$E\left(y_i \mid x_i\right) = \Pr\left(y_i = 1 \mid x_i\right) = \alpha_0 + \left(1 - \alpha_0 - \alpha_1\right) F\left(x_i'\beta\right).$$

The parameters of the binary response model with misclassification under the specified distribution of disturbance in the latent variable can be evaluated by non-linear least squares or by maximum likelihood. The non-linear least squares estimator can be set up to minimize the following sum of squares objective function to obtain the set of parameters (α_0, α_1, β):

$$\sum_{i=1}^{n} \left(y_i - a_0 - (1 - a_0 - a_1)F(x_i'b)\right)^2,$$

where standard parametric tests for the significance of the coefficients α_0 and α_1 can be used to measure the extent of misclassification in the model. The maximum likelihood estimator can be obtained by maximizing the log-likelihood function over the parameters (α_0, α_1, β):

$$\mathcal{L}\left(a_0, a_1, b\right) = \frac{1}{n} \sum_{i=1}^{n} y_i \ln\left(a_0 + \left(1 - a_0 - a_1\right) F(x_i'b)\right)$$
$$+ \left(1 - y_i\right) \ln\left(1 - a_0 - \left(1 - a_0 - a_1\right) F\left(x_i'b\right)\right).$$

The model of this type cannot be estimated as a ”classical” linear probability model where $F\left(x_i'b\right) = x_i'b$ because in that case, one cannot separately identify the parameters of the linear index $x_i'\beta$ and the factors α_0 and α_1. For identification of the parameters the authors require a monotonicity condition $\alpha_0 + \alpha_1 < 1$. In addition to this condition the authors impose a standard invertibility condition requiring that the matrix of regressors $E\left[xx'\right]$ is nonsingular, and that the distribution function $F\left(\cdot\right)$ of the disturbance in the latent variable is known.

Given the estimates of the model we can analyze the influence of misclassification on the parameters in the linear index driving the latent variable y^*. Specifically, define $\beta_E\left(\alpha_0, \alpha_1\right)$

to be the probability limits of the misspecified maximum likelihood estimates of β when the mismeasured y_i is used in place of the true $\tilde{y}_i$ in the log likelihood function, when the misclassification probabilities are α_0 and α_1 respectively. Therefore $\beta_E\left(\alpha_0,\,\alpha_1\right)$ is a function that characterizes the dependence of the estimate of the coefficient as a function of misclassification probabilities. In this case $\beta_E(0,0)=\beta$ is the coefficient in the model without misclassification. The marginal effects of misclassification can be derived as:

$$\begin{array}{l}\left|\frac{\partial\beta_E}{\partial\alpha_0}\right|_{\alpha_0=\alpha_1=0}=-\left[E\left(\frac{f(x'\beta)^2}{F(x'\beta)(1-F(x'\beta))}xx'\right)\right]^{-1}E\left(\frac{f(x'\beta)}{F(x'\beta)}x\right),\\ \left|\frac{\partial\beta_E}{\partial\alpha_1}\right|_{\alpha_0=\alpha_1=0}=\left[E\left(\frac{f(x'\beta)^2}{F(x'\beta)(1-F(x'\beta))}xx'\right)\right]^{-1}E\left(\frac{f(x'\beta)}{1-F(x'\beta)}x\right).\end{array}$$

Thus, the degree of inconsistency of the coefficients in the misclassified model with the true model will depend on the distribution of the disturbance and the regressor x. In general, the distributions with larger hazard functions will induce more bias in the estimation procedures which do not take into account misclassification.

If the marginal effects on binary response are of interest, they can be obtained by:

$$\frac{\partial\Pr(\tilde{y}=1|x)}{\partial x}=f\left(x'\beta\right)\beta,\qquad\text{for the true response}$$

$$\frac{\partial\Pr(y=1|x)}{\partial x}=\left(1-\alpha_0-\alpha_1\right)f\left(x'\beta\right)\beta,\quad\text{for the observed response.}$$

Thus, the difference between the true marginal effect and the marginal effect in the model with misclassification is increasing with the degree of misclassification, determined by the misclassification probabilities α_0 and α_1.

In many cases, the distribution of disturbances F is unknown. In that case the authors propose to use semiparametric estimation procedure. The authors establish the identification conditions for the semiparametric model with the flexible distribution of error in the latent variable. The two alternative sets of identification conditions include either monotonicity condition $\alpha_0+\alpha_1<1$ and the requirement that $F(\cdot)$ is strictly increasing, or the condition that $E\left(y\mid y^*\right)$ is increasing in y^* and the distribution function $F(\cdot)$ is strictly increasing.

The first condition is definitely stronger than the second one. However, it is similar to the assumptions of the parametric model and thus allows one to compare the performance of parametric and semiparametric model. In particular, we can run a specification test proposed in Horowitz and Hardle (1994) and if the parametric model is not rejected, then we can use it to improve the efficiency.

Given the established identification conditions, the authors set up the two-stage estimation procedure. In the first stage, they suggest to estimate the coefficient in the linear index β using maximum rank correlation (MRC) estimation based on Han (1987):

$$b_{MRC} = \underset{b}{\operatorname{argmax}} \sum_{i-1}^{n} Rank(x_i'b)y_i.$$

The constant term in b_{MRC} can not be identified, so the authors use a normalization of the index coefficient to estimate it. Moreover, the strong consistency and asymptotic normality of b_{MRC} have been proved (see Han (1987) and Sherman (1993)). The second stage makes use of the first stage estimated b_{MRC} and the observed dependent variables to obtain an estimation of the response function $G(\cdot)$ by isotonic regression and then to investigate the underlying misclassification mechanism. Define the estimated index value as $\hat{v}_i = x_i'b_{MRC}$, and the variables constructed in this way such that $\hat{v}_1 \leq \hat{v}_2 \leq \cdots \leq \hat{v}_n$ The resulting response function $\hat{G}$ is a so-called isotonic function - it is non-decreasing on the set of n index values. To find this function for $v \in (\hat{v}_i, \hat{v}_{i+1})$ we find values $\hat{G}$ minimizing:

$$\sum_{i=1}^{n} (y_i - \hat{G}(\hat{v}_i))^2$$

over the set of isotonic functions; for $v < \hat{v}_1$, $\hat{G}(v) = 0$; for $v > \hat{v}_n$, $\hat{G}(v) = 1$. It can be shown that $\hat{G}$ is $\sqrt[3]{n}$-consistent. Moreover the asymptotic distribution of the point estimates of the response function can be described as:

$$\frac{n^{\frac{1}{3}}(\hat{G}(v) - G(v))}{\frac{1}{2}G(v)(1 - G(v)\frac{g(v)}{h(v)})^{\frac{1}{3}}} \to 2Z,$$

where the random variable Z is the last time where two-sided Brownian motion minus the parabola u^2 reaches its maximum, $g(\cdot)$ is the derivative of the response function, and $h(\cdot)$ is the density of the linear index in the latent variable. The two sided Brownian motion is defined as a stochastic process Z_t constructed from two independent Brownian motions B_t^+ and B_t^- such that if the index $t > 0$ then $Z_t = B_t^+$ and if $t < 0$ then $Z_t = B_{-t}^-$. The distribution of Z can be written as:

$$f_Z(u) = \frac{1}{2}s(u)s(-u), \text{ for } u \in \mathbb{R},$$

where the function $s(\cdot)$ has a Fourier transform:

$$\hat{s}(w) = \frac{2^{1/3}}{Ai(2^{-1/3}wi)}.$$

In this expression $Ai(\cdot)$ is the Airy function which is defined as a bounded solution of the differential equation $x'' - tx = 0$.

The convergence of the constructed semiparametric estimator is slower than the convergence of estimator from the parametric model. An attractive feature of the semiparametric approach is that it allows one to estimate the parameters β in the linear index under weaker assumptions. Semiparametric model can be useful even in the case when the we know that the structure of the data generating process is the same as in the parametric model. Specifically, if $g(\cdot)$ is the derivative of the conditional expectation of y given x then the marginal effect can be represented as:

$$\frac{\partial \Pr(\widetilde{y} = 1 \mid x)}{\partial x} = \frac{g(x'\beta)\beta}{1 - \alpha_0 - \alpha_1}.$$

The apparent lower bound for the marginal effect is achieved in the absence of misclassification when the marginal effect is equal to $g(x_i'\beta)\beta$. In case when some consistent estimates of the misclassification are available one can correct the marginal effect for misclassification. In principal, these probabilities can be inferred from the asymptotic behavior of the conditional expectation $E[y \mid x] = G(x'\beta)$. Specifically, according to the expression for the conditional expectation in terms of the cumulative distribution of the disturbance in the latent variable y^* the limit behavior gives us expressions for the misclassification probabilties $\lim\limits_{z \to -\infty} G(z) = \alpha_0$ and $\lim\limits_{z \to +\infty} G(z) = 1 - \alpha_1$. Out-of sample fit of semiparametric estimates can be poor and in general we cannot use them for precise predictions. However, using, for instance, the results in Horowitz and Manski (1995) we can use the results of semiparametric analysis to form upper bounds for the misclassification probabilities. These bounds will provide an upper bound for the estimated marginal effect.

In Hausman, Abrevaya, and Scott-Morton (1998) the authors apply their semi-parametric technique to study a model of job change using data from the Current Population Survey

(CPS) and the Panel Study of Income Dynamics. Using these two datasets the authors can evaluate the probabilities of job change over certain periods of time. According to the authors, the questions about job tenure are not always understood by the respondents and, thus, the survey data contain a certain amount of misclassification error connected with the wrong responses of individuals. Using the methodology of the paper, it is possible to correct the bias in the estimates of the probabilities of job change connected with the misclassification errors in the data. As the authors report, the construction of the job tenure variable in a standard way leads to a substantial bias in the estimates, while the methods provided by the authors allow them to correct the bias due to misclassification.

4.1.2 Misclassification of discrete regressors using IVs

Recently **Mahajan (2005)** studies a nonparametric regression model where one of the true regressors is a binary variable:

$$E\left\{y - g\left(x^{*}, z\right) \mid \left(x^{*}, z\right)\right\} = 0.$$

In this model the variable x^{*} is binary and z is continuous. The true binary variable x^{*} is unobserved and the econometrician observes a potentially misreported value x instead of x^{*}. Mahajan (2005) assumes that in addition, another random variable v is observed. The variable v takes at least two values v_1 and v_2. The author mentions that the variable v plays the role of an exclusion restriction in the standard instrumental variable estimation. Mahajan (2005) imposes the following assumptions on the model.

Assumption 1 *The regression function $g(x^{*}, z)$ is identified given the knowledge of the population distribution of $\{y,\, x^{*},\, z\}$.*

Assumption 1 implies that the regression function is identifiable in the absence of the measurement error. This means that the incomplete identification of the model with the measurement errors is possible only when the model is identified without measurement errors.

The second assumption restricts the extent of possible misclassification so that the observed signal is not dominated by misclassification noise. Denote

$$\alpha_0(z) = \Pr\left(x = 1 \mid x^{*} = 0,\, z\right),\ \alpha_1(z) = \Pr\left(x = 0 \mid x^{*} = 1,\, z\right)$$

as the probabilities of misclassification.

Assumption 2 $\alpha_0(z) + \alpha_1(z) < 1$.

This assumption suggests that the observed proxy x for the dependent variable is positively correlated with the unobserved true variable x^* and, thus, allows one to retrieve some information about x^*. This assumption can be substituted by the assumption that the sign of the correlation of the proxy variable and the unobserved true variable is known. One can see, that the latter assumption will be equivalent to Assumption 2 if one substitutes x by $-x$ in case of a negative correlation.

The third assumption declares independence of the proxy variable from the binary variable v conditional on the true variable x^* and a continuous variable z.

Assumption 3 $x \perp v \mid (x^*, z)$.

This assumption is similar to the identification assumptions in the previous literature and it is important for the point identification in the model with a mismeasured regressor.

The next assumption requires that the conditional probability of the true regressor actually depends on the binary instrument v.

Assumption 4 *There exist* $v_1 \neq v_2$ *such that* $Pr(x^* = 1 \mid z, v_1) \neq Pr(x^* = 1 \mid z, v_2)$.

This assumption implies that the instrumental variable is informative about the unobserved regressor x^* even given the other covariates and, thus, suggests that the instrument v is informative for x^*

The last assumption requires that the unobserved regressor x^* is relevant for the conditional expectation under consideration.

Assumption 5 $g(1, z) \neq g(0, z)$.

The author mentions that this assumption is potentially testable because it implies that the expectation of the variable y conditional on observable x and z should be different for $x = 0$ and $x = 1$.

An important result of the author is that under assumptions 1-5 both the value of the regression function $g(x^*, z)$ and the values of the misclassification probabilities are identified. Moreover, if assumptions 1-5 are formulated for almost all z on its support, then the entire regression function and the miclassification probabilities as functions of z are identified.

To see this, denote $\eta_2(z,v) = \Pr(x = 1 \mid z,\, v)$ and $\eta_2^*(z,v) = \Pr(x^* = 1 \mid z,\, v)$. Note that $\eta_2(z,v)$ is observable and note the following relations:

$$\begin{cases} E(x|z,v) \equiv \eta_2(z,v) = (1-\eta_1(z))\,\eta_2^*(z,v) + \eta_0(z)(1-\eta_2^*(z,v)), \\ E(y|z,v) = g(1,z)\,\eta_2^*(z,v) + g(0,z)(1-\eta_2^*(z,v)), \\ E(yx|z,v) = g(1,z)(1-\eta_1(z))\,\eta_2^*(z,v) + g(0,z)\,\eta_0(z)(1-\eta_2^*(z,v)). \end{cases} \tag{30}$$

Suppose v takes n_v values. For each z, $\eta_0(z)$, $\eta_1(z)$, $g(0,z)$, $g(1,z)$ and $\eta_2^*(z,v)$ are unknown. There are $4+n_v$ parameters, and $3n_v$ equations. Therefore as long as $n_v \geq 2$, all the parameters can possibly be identified. Intuitively, if $\eta_2^*(z,v)$ is known, the second moment condition $E(y|z,v)$ identifies $g(1,z)$ and $g(0,z)$. Information from the other moment conditions also allows one to identify both $\eta_1(z)$ and $\eta_0(z)$.

A constructive proof is given in Mahajan (2005) using the above three moment conditions. Rearranging the first moment condition, one obtains

$$\eta_2^*(z,v) = \frac{\eta_2(z,v) - \eta_0(z)}{1-\eta_0(z)-\eta_1(z)}.$$

Substituting this into the next two moment conditions, one can write

$$\begin{aligned} E(y|z,v) =& g(0,z) + (g(1,z)-g(0,z))\frac{\eta_2(z,v)-\eta_0(z)}{1-\eta_0(z)-\eta_1(z)} \\ =& g(0,z) - \frac{(g(1,z)-g(0,z))\,\eta_0(z)}{1-\eta_0(z)-\eta_1(z)} + \frac{g(1,z)-g(0,z)}{1-\eta_0(z)-\eta_1(z)}\eta_2(z,v) \\ E(yx|z,v) =& g(0,z)\,\eta_0(z) - [g(1,z)(1-\eta_1(z)) - g(0,z)\,\eta_0(z)]\frac{\eta_0(z)}{1-\eta_0(z)-\eta_1(z)} \\ &+ \frac{[g(1,z)(1-\eta_1(z)) - g(0,z)\,\eta_0(z)]}{1-\eta_0(z)-\eta_1(z)}\eta_2(z,v) \\ =& -\frac{(g(1,z)-g(1,z))\,\eta_0(z)(1-\eta_1(z))}{1-\eta_0(z)-\eta_1(z)} \\ &+ \frac{[g(1,z)(1-\eta_1(z)) - g(0,z)\,\eta_0(z)]}{1-\eta_0(z)-\eta_1(z)}\eta_2(z,v). \end{aligned}$$

Mahajan (2005) suggests that if one runs a regression of $E\left(y|z,v\right)$ on $\eta_2\left(z,v\right)$ and a regression of $E\left(yx|z,v\right)$ on $\eta_2\left(z,v\right)$, then one can recover the intercepts and the slope coefficients:

$$\begin{aligned}
a =& g\left(0,z\right) - \frac{\left(g\left(1,z\right)-g\left(0,z\right)\right)\eta_0\left(z\right)}{1-\eta_0\left(z\right)-\eta_1\left(z\right)} \\
b =& \frac{g\left(1,z\right)-g\left(0,z\right)}{1-\eta_0\left(z\right)-\eta_1\left(z\right)} \\
c =& g\left(0,z\right)\eta_0\left(z\right) - \left[g\left(1,z\right)\left(1-\eta_1\left(z\right)\right)-m\left(0,z\right)\eta_0\left(z\right)\right]\frac{\eta_0\left(z\right)}{1-\eta_0\left(z\right)-\eta_1\left(z\right)} \\
=& -\frac{\left(g\left(1,z\right)-g\left(1,z\right)\right)\eta_0\left(z\right)\left(1-\eta_1\left(z\right)\right)}{1-\eta_0\left(z\right)-\eta_1\left(z\right)} \\
d =& \frac{\left[g\left(1,z\right)\left(1-\eta_1\left(z\right)\right)-g\left(0,z\right)\eta_0\left(z\right)\right]}{1-\eta_0\left(z\right)-\eta_1\left(z\right)}.
\end{aligned}$$

Therefore one can write

$$a = \quad m\left(0,z\right)-\eta_0\left(z\right)b \tag{31}$$

$$c = \quad m\left(0,z\right)\eta_0\left(z\right)-d\eta_0\left(z\right) \tag{32}$$

and

$$c = -b\left(1-\eta_1\left(z\right)\right)\eta_0\left(z\right). \tag{33}$$

Equations (31) can be used to concentrate out $m\left(0,z\right)$. One can then substitute it into (32) and make use of (33) to write

$$\left(a+\eta_0\left(z\right)b\right)\eta_0\left(z\right)-d\eta_0\left(z\right) = -b\left(1-\eta_1\left(z\right)\right)\eta_0\left(z\right).$$

Then one can factor out $\eta_0\left(z\right)$ and rearrange:

$$1-\eta_1\left(z\right)+\eta_0\left(z\right) = \frac{d-a}{b}. \tag{34}$$

Now we have two equations (33) and (34) in two unknowns $1-\eta_1\left(z\right)$ and $\eta_0\left(z\right)$. Obviously the solutions to this quadratic system of equation is only unique up to an exchange between $1-\eta_1\left(z\right)$ and $\eta_0\left(z\right)$. However, Assumption 2 rules out one of these two possibilities and allows for point identification; hence Mahajan (2005) demonstrates that the model is identified.

Mahajan (2005) further develops his identification strategy into a nonparametric estimator, and also provides a semiparametric estimator for a single index model. Specifically, for a known index function $\tau(\cdot;\theta)$ the author considers a model represented by the moment equation:

$$E\left[y - g\left(\tau\left(x^{*}, z; \theta\right)\right) \mid x^{*},\, z\right] = 0.$$

Under the assumption that if the unobserved x^{*} is known, then the proxy x does not provide additional information about it, the moment condition can be transformed to:

$$E\left[y - g\left(\tau\left(x^{*}, z; \theta\right)\right) \mid x^{*},\, z,\, x,\, v\right] = 0.$$

In this case the additional assumptions for identification of the semiparametric model are the following.

Assumption 6 *The parameter θ_0 and the function $g(\cdot)$ in the semiparametric moment condition model are identified given the distribution of $\{y,\, x^{*},\, z\}$.*

The next assumption declares the relevance of the unobserved variable x^{*} for the values of the linear index.

Assumption 7 $\tau\left(1,\, z,\, \theta_0\right) \neq \tau\left(0, z, \theta_0\right)$ *almost everywhere on the support of z*

Assumptions 6 and 7, in addition to assumptions 2-4, for almost all z on its support will allow for both the parameter vector θ_0 and the misclassification rates to be identified. This is a corollary from the general nonparametric identification argument of the author.

The semiparametric model can be further simplified to a fully parametric model such as a parametric binary choice model. In this case the additional identification assumptions reduce to the invertability of the matrix of independent variables and non-zero coefficient for the unobserved binary variable.

The estimation procedure suggested by the author follows the identification argument of the model which we briefly described above. Specifically, one estimate the system of moments (30) by kernel smoothing and then solve it for the three unknown functions. The author also provided additional conditions for appropriate asymptotic behavior of the obtained estimates. These assumptions include uniform boundedness of the products of the conditional density of the observed continuous variable z and second moments of the expressions in the moment conditions, uniform boundedness and continuity of the distributions and moments under analysis and standard assumptions about the kernel function. In this case both the empirical moments and the estimates of the regression function are asymptotically normal with the non-parametric convergence rate $\sqrt{nh_n}$, where h_n is the kernel bandwidth parameter.

Mahajan (2005) suggests a constructive and simple test for misclassification. The idea of the test is that the instrument v is relevant for the outcome variable y only in the case when there is misclassification (because the information about the true regressor x^* is sufficient for locating the conditional expectation of the outcome y). Mahajan (2005) first proves that both misclassification probabilities are zero $\eta_0(z) = \eta_1(z) = 0$ if and only if the instrument v is not relevant, so that

$$E\left(y \mid x,\, z,\, v\right) = E\left(y \mid x,\, z\right).$$

In this case the test for misclassification can be conducted as a test for the equality of the outlined conditional expectations, both of which can be estimated non-parametrically. The test statistic is constructed as a difference between the two estimated expectations and it should converge to the normal distribution with zero mean at a non-parametric rate. The efficiency of the test can be increased by switching to a semi-parametric model, while in the fully parametric model the test reduces to the standard test for the exclusion restriction.

An alternative approach to identification and estimation of the model with a misclassified binary regression is considered in the paper by **Hu (2006)**. The author looks at a general problem of identification of the joint density:

$$f_{y|x^*,z}\left(y \mid x^*,\, z\right).$$

Here y is a one-dimensional random variable, x^* is the unobserved discrete part of regressor and z is the observed part regressor. One can observe a proxy for the unobserved regressor x^* - a binary variable x and an instrument v. It is assumed that the variables x, x^* and v have a common discrete support $\{1, 2, \ldots, k\}$. The assumptions of the author are close to those in Mahajan (2005). Hu (2006) assumes that the value of the unobserved regressor

x^* provides sufficient information about the outcome variable y so that if the value of x^* is known, then the information about the proxy variable x and the instrument v is redundant.

Assumption 8 $f_{y|x^*,x,z,v}\left(y \mid x^*,\, x,\, z,\, v\right) = f_{y|x^*,z}\left(y \mid x^*,\, z\right).$

Similar to Mahajan (2005), assumption 8 states that the misclassification error in the proxy variable x is independent of the dependent variable y conditional on the true regressor $(x^*,\, z)$, and is also conditional independent of the instrument v.

The next assumption, same as assumption 3, requires that the misclassification error in the proxy variable x is independent of the instrument v conditional on the true regressor (x^*, z). It suggests that the information about the value of the unobserved binary variable x^* and the observed part of the regressor z is sufficient to determine the distribution of the proxy variable. A particular case when this assumption will hold is the case of the classical measurement error.

Assumption 9 $f_{x|x^*,z,v}\left(x \mid x^*,\, z,\, v\right) = f_{x|x^*,z}\left(x \mid x^*,\, z\right).$

The further analysis of the author suggests that one can form a system of equations relating the observed distributions to the unobserved distributions and then find the unobserved distributions by simple matrix inversion. This approach generalizes that of Mahajan (2005) to the case of multiple values of the misclassified regressor.

To conduct the analysis of identification and develop the estimation procedure for the model, it is convenient to define the following matrices. Denote

$$\begin{aligned}
&\mathbf{F}_{yx|vz} = \left(f_{yx|vz}(y,\, i \mid j,\, z)\right)_{i,j=1}^{k}, \quad \mathbf{F}_{x^*|vz} = \left(f_{x^*|vz}(i \mid j,\, z)\right)_{i,j=1}^{k},\\
&\mathbf{F}_{x|x^*z} = \left(f_{x|x^*z}(i \mid j,\, z)\right)_{i,j=1}^{k}, \quad \mathbf{F}_{y|x^*z} = diag\left\{f_{y|x^*z}(y \mid i,\, z),\ i = 1, \ldots, k\right\},\\
&\mathbf{F}_{y|zv} = \left(f_{y|vz}(y \mid i,\, z)\right)_{i=1}^{k}.
\end{aligned}$$

Under Assumptions 8 and 9, the expressions for conditional distributions in the matrix form are:

$$\begin{aligned}
\mathbf{F}_{yx|vz} &= \mathbf{F}_{x^*|vz}\,\mathbf{F}_{y|x^*z}\,\mathbf{F}_{x|x^*z},\\
\mathbf{F}_{x|vz} &= \mathbf{F}_{x^*|vz}\,\mathbf{F}_{x|x^*z}.
\end{aligned} \tag{35}$$

The additional equation comes from the definition of conditional density and takes the form:

$$\mathbf{F}_{y|vz} = \mathbf{F}_{x^*|vz}\,\mathbf{F}_{y|x^*z}\,\mathbf{1}, \tag{36}$$

where $\mathbf{1}$ is a $k \times 1$ vector of ones.

To resolve the system of equations (35) and (36), the author adds the following assumption which generalizes assumption 5:

Assumption 10 *Rank* $\left(\mathbf{F}_{x^*|v,z}\right) = k$.

In addition, assuming that the matrix $\mathbf{F}_{x|x^*z}$ is non-singular, we can form a system of equations for the unknown $k(k+1)$ elements of $\mathbf{F}_{y|x^*z}$ and $\mathbf{F}_{x|x^*z}$ for every possible y and z in the form:

$$\begin{aligned} &\mathbf{F}_{x|x^*z}\,\mathbf{F}_{x|vz}^{-1}\,\mathbf{F}_{yx|vz}\,\mathbf{F}_{x|x^*z}^{-1} = \mathbf{F}_{y|x^*z}, \\ &\mathbf{F}_{x|x^*z}\,\mathbf{1} = \mathbf{1}. \end{aligned} \tag{37}$$

Denote $\mathbf{A} = \mathbf{F}_{x|vz}^{-1}\,\mathbf{F}_{yx|vz}$ which is constructed from the matrices of observable distributions. Since the matrix $\mathbf{F}_{y|x^*z}$ is diagonal and is expressed in a "sandwich" form in terms of the matrix $\mathbf{A}$, $\mathbf{F}_{y|x^*z}$ and $\mathbf{A}$ have the same eigenvalues. Even though the matrix $\mathbf{A}$ can be reconstructed from the data, without additional assumptions it will be impossible to map its eigenvalues to the elements of the matrix $\mathbf{F}_{y|x^*z}$. To make this mapping the author imposes additional restrictions on the distributions of the model. He first requires that there is a function $\gamma(\cdot)$ such that the expectation $E\left[\gamma(y) \mid x^* = i, z\right] \neq E\left[\gamma(y) \mid x^* = j, z\right]$ for all $i \neq j$. Additionally the author requires that the conditional distribution $f_{y|x^*z}$ is strictly monotone in x^* for every y and z. Under these additional restrictions we will be able to associate the values of the density with the ordered eigenvalues of the matrix $\mathbf{A}$. Furthermore, the matrix of eigenvectors of the matrix $\mathbf{A}$ can be associated with the matrix of misclassification probabilities $\mathbf{F}_{x|x^*z}$.

The author notes that an equivalent identification assumption is to assume that there exists a function $\omega(\cdot)$ such that the conditional expectation $E\left[\omega(y) \mid x^*, z\right]$ is strictly increasing in x^*. Such an assumption does not imply any restrictions for the matrix $\mathbf{F}_{x|x^*z}$ *per se* and thus is quite flexible with respect to the distribution of measurement errors. As an alternative, to identify the distribution $\mathbf{F}_{x|x^*z}$ we can impose restrictions on this distribution directly, requiring its monotonicity or, alternatively, the domination of the upper-triangular components of this matrix of misclassification probabilities.

The estimation strategy suggested in Hu (2006) is suited for the case of semiparametric specification, when the outcome variable is described by a moment relation:

$$E\left(y \mid x^{*},\, z\right)=m^{*}\left(x^{*},\, z\,;\, \theta_{0}\right).$$

In this case the unknown distribution $f_{x^{*}|x,z}$ is obtained from the eigenvalue decomposition suggested in the proof of identification and the moment $m^{*}(\cdot)$ can be transformed to a function of observable variables y, x, z and v to form a GMM-type objective. Given a set of smoothness and uniform boundedness assumptions on the moments and distributions of the model, the author proves that the estimate of the parameter θ_0 is asymptotically normal with a parametric convergence rate $\sqrt{n}$.

As an application, Hu (2006) analyzes the impact of education on women's fertility. The author specifies the moment condition in the exponential form. To characterize the distribution of the dependent variable, the author uses a quasi-maximum likelihood estimator based on the Poisson distribution. The author uses the data from the Current Population Survey and first estimates the parameters of the moment equation without taking into account possible misclassification errors in the regressor - the level of education. The author then compares the performance of the method in the paper with the performance of the standard quasi-maximum likelihood estimator. It appears that the standard QMLE estimates of the semi-elasticity of the number of children with respect to education are biased towards zero. This means that if the mathodology does not take into account the measurement errors the effects of the policy changes might be underevaluated. The estimates of the semi-elasticity obtained using the method of the paper are almost twice the estimates obtained using the QMLE. As a specification test the author uses a Hausman-type test to verify the presence of the measurement error in the data. The test suggests that the hypothesis of the absence of the measurement errors in the data is rejected.

In a related paper, **Lewbel (2006)** considers a model with a binary regressor that can be mismeasured. His model is similar to a model of average treatment effect when the treatment is observed with an error. Namely, the author considers estimation of the non-parametric regression function $E\left(Y \mid z, v, x^{*}\right)$ where the true binary regressor x^{*} is mismeasured as x. Y is the observed treatment outcome, and (z, v) is the set of instruments.

A specific object of interest of the author is the average treatment effect

$$\tau^* (z,\, v) = E\left(Y \mid z,\, v,\, x^* = 1\right) - E\left(Y \mid z,\, v,\, x^* = 0\right),$$

so that the regression model given the true treatment dummy can be written as:

$$E\left(Y \mid z,\, v,\, x^*\right) = E\left(Y \mid z,\, v,\, x^* = 0\right) + \tau^* (z,\, v)\, x^*.$$

If we define y_0 to be the variable corresponding to the treatment outcome when $x^* = 0$ and y_1 as the outcome when $x^* = 1$, then the conditional average treatment effect is defined as:

$$\widetilde{\tau}(z, v) = E\left[y_0 - y_1 \mid z, v\right].$$

To relate the average treatment effect to the conditional treatment effect, the author imposes two restrictions on the distribution of the outcomes. First, the author assumes that:

$$E\left(Y \mid z,\, v,\, x^*,\, x\right) = E\left(Y \mid z,\, v,\, x^*\right),$$

which means that given the true treatment dummy, the mismeasured treatment dummy does not add information to the conditional expectation. The second assumption is similar to the assumption in Mahajan (2005), which limits the extent of misclassificiation. More specifically, the author assumes that:

$$\Pr\left(x = 0 \mid z,\, v,\, x^* = 1\right) + \Pr\left(x = 1 \mid z,\, v,\, x^* = 0\right) < 1.$$

In addition to this, the author assumes that the treatment probability is positive but not all of the outcomes are treated:

$$0 < E\left(x^* \mid z,\, v\right) < 1.$$

These two assumptions allow the author to prove the result that if $\tau\left(z, v\right)$ is the observed treatment effect (estimated from the mismeasured treatment dummy), then relation between the true treatment effect and the observed treatment effect is given by:

$$\tau(z, v) = m(z, v)\tau^*(z, v).$$

The function $m(z,v)$ can be expressed through the observed treatment probability $P\left(x|z,v\right)$ and the unobserved probabilities of observed treatment dummies given the true treatment dummies $P\left(x|x^{*},z,v\right)$. Denoting the observed probability of treatment as $r(z,v) = E\left(x \mid z,v\right)$, we can express $m(z,v)$ as:

$$\begin{aligned} m(z,v) &= \left(1-\Pr\left(x=0 \mid z,\, v,\, x^{*}=1\right)-\Pr\left(x=1 \mid z,\, v,\, x^{*}=0\right)\right)^{-1} \\ &+ \Big(1-\left[1-\Pr\left(x=0 \mid z,\, v,\, x^{*}=1\right)\right]\Pr\left(x=1 \mid z,\, v,\, x^{*}=0\right)r(z,v)^{-1} \\ &- \left[1-\Pr\left(x=1 \mid z,\, v,\, x^{*}=0\right)\right]\Pr\left(x=0 \mid z,\, v,\, x^{*}=1\right)\left(1-r(z,v)\right)^{-1}\Big). \end{aligned}$$

Under the imposed assumption this function is bounded by $0 < m(z,v) \leq 1$. This means that identification of the true treatment effect requires additional restrictions on the probability distributions of the observed variables. For this reason the author imposes two additional assumptions.

The first assumption requires that for some subset of the support of (z,v), we can fix z and the variation in v does not lead to the changes in the conditional probability of the observed treatment and the true treatment effect, but changes the probability of the true treatment dummy. More formally, there exists $\mathcal{A} \in \operatorname{supp}\left(z,v\right)$ such that for all $((z,v)\ (z,v')) \in \mathcal{A}$ where $v' \neq v$ we have:

$$\begin{aligned} &\Pr\left(x=1 \mid z,\, v,\, x^{*}=0\right) = \Pr\left(x=1 \mid z,\, v',\, x^{*}=0\right), \\ &\Pr\left(x=0 \mid z,\, v,\, x^{*}=1\right) = \Pr\left(x=0 \mid z,\, v',\, x^{*}=1\right), \\ &\tau^{*}(z,v) = \tau^{*}(z,v'), \end{aligned}$$

but $r^{*}(z,v) \neq r^{*}(z,v')$.

The next assumption imposes a "sufficient variation" restriction on the conditional probability of the treatment outcome. Specifically the author assumes that it is possible to find three elements in the support of (z,v) with $v_0,\ v_1,\ v_2$, and the same component z, such that:

$$\left(\tfrac{\tau(v_0,z)}{r(v_1,z)}-\tfrac{\tau(v_1,z)}{r(v_0,z)}\right)\left(\tfrac{\tau(v_0,z)}{1-r(v_2,z)}-\tfrac{\tau(v_2,z)}{1-r(v_0,z)}\right) \neq \left(\tfrac{\tau(v_0,z)}{r(v_2,z)}-\tfrac{\tau(v_2,z)}{r(v_0,z)}\right)\left(\tfrac{\tau(v_0,z)}{1-r(v_1,z)}-\tfrac{\tau(v_1,z)}{1-r(v_0,z)}\right).$$

Under these assumptions, the true treatment effect $\tau^{*}\left(z,v\right)$, the misclassification probabilities $P\left(x|z,v,x^{*}\right)$ and the probability of treatment $r^{*}\left(z,v\right)$ are all identified.

In addition to this, if the restriction on the misclassification probabilities is substituted by

$$\Pr\left(x=0 \mid z,\, v,\, x^{*}=1\right)+\Pr\left(x=1 \mid z,\, v,\, x^{*}=0\right) \neq 1,$$

then the treatment effect is identified up to a change of the sign.

Lewbel (2006) suggests a GMM estimation method when the support of the instrument v is discrete with K elements, $\{v_k, k=1,\ldots,K\}$. The estimation is based on two moments. the first moment equation is expressing the unconditional probability of the observed treatment dummy in terms of the probabilities of mismeasurement:

$$\begin{aligned} E\left(\{\Pr\left(x=1 \mid z,\, v_k,\, x^{*}=0\right)+\left[1-\Pr\left(x=0 \mid z,\, v_k,\, x^{*}=1\right)\right.\right. \\ \left.\left. -\Pr\left(x=1 \mid z,\, v_k,\, x^{*}=0\right)\right] r^{*}\left(v_k, z\right)-x\} \mid z,\, v_k\right)=0. \end{aligned}$$

The second moment equation makes use of the established relationship between the observed and the true treatment effect:

$$\begin{aligned} & E\,(\ \tau^{*}(z,v)\mathbf{1}\{v=v_k\} \\ & +\frac{yx-\left[1-\Pr\left(x=1 \mid z,\, v,\, x^{*}=0\right)\right] r^{*}(z,v_k)\tau^{*}(z,v)\mathbf{1}\{v=v_k\}}{\Pr\left(x=0 \mid z,\, v,\, x^{*}=1\right)+\left(1-\Pr\left(x=0 \mid z,\, v,\, x^{*}=1\right)-\Pr\left(x=1 \mid z,\, v,\, x^{*}=0\right) r^{*}(z,v_k)\right)} \\ & +\left[y(1-x)+\left[1-\Pr\left(x=0 \mid z,\, v,\, x^{*}=1\right)\right] r^{*}(z,v_k)\tau^{*}(z,v)\mathbf{1}\{v=v_k\}\right]\left[1-\{\Pr\left(x=0 \mid z,\, v,\, x^{*}=1\right)\right. \\ & \left.+\ \left(1-\Pr\left(x=0 \mid z,\, v,\, x^{*}=1\right)-\Pr\left(x=1 \mid z,\, v,\, x^{*}=0\right) r^{*}(z,v_k)\right)\}\right]^{-1} \mid z,\, v\ )=0. \end{aligned}$$

The author then applies the GMM procedure to solve for the unknown functions assuming a parametric form for the unknown probability distributions and semi-parametric specification for the distribution of the covariates z.

Lewbel (2006) then apply his identification and estimation procedure to study the effect of having a college degree on earnings, given that the completion of college may be misreported. The author uses data from the National Longitudinal Survey of the High School class of 1972 to obtain information about wages and data from the Post-secondary Education Transcript Survey to obtain information about transcripts from which one can infer the completion of a college degree. As an instrument with a discrete support, the author uses the rank data about the distance from the respondent's high school to the closest four-year college. The author uses experience and demographic variables as additional covariates. To simplify the analysis the author suggests a parametric specification for the probabilities of misreporting, the probability of the true binary regressor (indicating the college degree), and the treatment effect, which is assumed to depend linearly on covariates. Then the parameters of interest are estimated by GMM. The author finds that misclassification has a large impact on the obtained estimates with a significant downward bias: "naive" estimation gives an impact of 11% from the college degree, while the GMM estimates suggest an impact of 38%.

4.1.3 Misclassification of discrete regressors using two samples

Recently **Chen and Hu (2006)** consider identification and estimation of general nonlinear models with nonclassical measurement errors using two samples, where both samples contains measurement errors and neither sample contains an accurate observation of the truth nor the presence of instrumental variables. We illustrate their identification strategy by describing a special case in which the key variables in the model are 0-1 dichotomous. Suppose that we are interested in the effect of the true college education level X^* on the labor supply Y with the marital status W^u and the gender W^v as covariates. This effect would be identified if we could identify the joint density $f_{X^*,W^u,W^v,Y}$. We assume X^*, W^u, and W^v are all 0-1 dichotomous. The true education level X^* is unobserved and is subject to measurement errors. (W^u, W^v) are accurately measured and observed in both the primary sample and the auxiliary sample, and Y is only observed in the primary sample. The primary sample is a random sample from (X, W^u, W^v, Y), where X is a mismeasured X^*. In the auxiliary sample, we observe (X_a, W_a^u, W_a^v), in which the observed X_a is a proxy of a latent education level X_a^*, W_a^u is the marital status, and W_a^v is the gender. In this illustration, we use italic letters to highlight all the assumptions imposed by Chen and Hu (2006) for the nonparametric identification of $f_{X^*,W^u,W^v,Y}$.

The authors assume that *the measurement error in X is independent of all other variables in the model conditional on the true value X^*, i.e., $f_{X|X^*,W^u,W^v,Y} = f_{X|X^*}$.* Under this assumption, the probability distribution of the observables equals

$$f_{X,W^u,W^v,Y}\left(x,u,v,y\right) = \sum_{x^*=0,1} f_{X|X^*}(x|x^*) f_{X^*,W^u,W^v,Y}(x^*,u,v,y) \quad \text{for all } x,u,v,y. \tag{38}$$

Define the matrix representations of $f_{X|X^*}$ as follows:

$$L_{X|X^*} = \begin{pmatrix} f_{X|X^*}(0|0) & f_{X|X^*}(0|1) \\ f_{X|X^*}(1|0) & f_{X|X^*}(1|1) \end{pmatrix}.$$

Notice that the matrix $L_{X|X^*}$ contains the same information as the conditional density $f_{X|X^*}$. Equation (38) then implies for all u, v, y

$$\begin{pmatrix} f_{X,W^u,W^v,Y}(0,u,v,y) \\ f_{X,W^u,W^v,Y}(1,u,v,y) \end{pmatrix} = L_{X|X^*} \times \begin{pmatrix} f_{X^*,W^u,W^v,Y}(0,u,v,y) \\ f_{X^*,W^u,W^v,Y}(1,u,v,y) \end{pmatrix}. \tag{39}$$

Equation (39) implies that the density $f_{X^*,W^u,W^v,Y}$ would be identified provided that $L_{X|X^*}$ would be identifiable and invertible. Moreover, equation (38) implies, for the subsamples of males ($W^v = 1$) and of females ($W^v = 0$)

$$\begin{aligned} f_{X,W^u|W^v=j}(x,u) &= \sum_{x^*=0,1} f_{X|X^*,W^u,W^v=j}(x|x^*,u)\, f_{W^u|X^*,W^v=j}(u|x^*) f_{X^*|W^v=j}(x^*). \\ &= \sum_{x^*=0,1} f_{X|X^*}(x|x^*)\, f_{W^u|X^*,W^v=j}(u|x^*) f_{X^*|W^v=j}(x^*), \end{aligned} \tag{40}$$

in which $f_{X,W^u|W^v=j}(x,u) \equiv f_{X,W^u|W^v}(x,u|j)$ and $j = 0, 1$.

The authors assume that, in the auxiliary sample *the measurement error in X_a satisfies the same conditional independence assumption as that in X, i.e., $f_{X_a|X_a^*,W_a^u,W_a^v} = f_{X_a|X_a^*}$.* Furthermore, they link the two samples by a stable assumption that *the distribution of the marital status conditional on the true education level and gender is the same in the two samples, i.e., $f_{W_a^u|X_a^*,W_a^v=j}(u|x^*) = f_{W^u|X^*,W^v=j}(u|x^*)$ for all u, j, x^*.* Therefore, one has for the subsamples of males ($W_a^v = 1$) and of females ($W_a^v = 0$):

$$\begin{aligned} f_{X_a,W_a^u|W_a^v=j}(x,u) &= \sum_{x^*=0,1} f_{X_a|X_a^*,W_a^u,W_a^v=j}(x|x^*,u)\, f_{W_a^u|X_a^*,W_a^v=j}(u|x^*) f_{X_a^*|W_a^v=j}(x^*) \\ &= \sum_{x^*=0,1} f_{X_a|X_a^*}(x|x^*)\, f_{W^u|X^*,W^v=j}(u|x^*) f_{X_a^*|W_a^v=j}(x^*). \end{aligned} \tag{41}$$

Define the matrix representations of relevant densities for the subsamples of males

$(W^v = 1)$ and of females $(W^v = 0)$ in the primary sample as follows: for $j = 0, 1$,

$$\begin{aligned} L_{X,W^u|W^v=j} &= \begin{pmatrix} f_{X,W^u|W^v=j}(0,0) & f_{X,W^u|W^v=j}(0,1) \\ f_{X,W^u|W^v=j}(1,0) & f_{X,W^u|W^v=j}(1,1) \end{pmatrix} \\ L_{W^u|X^*,W^v=j} &= \begin{pmatrix} f_{W^u|X^*,W^v=j}(0|0) & f_{W^u|X^*,W^v=j}(0|1) \\ f_{W^u|X^*,W^v=j}(1|0) & f_{W^u|X^*,W^v=j}(1|1) \end{pmatrix}^T \\ L_{X^*|W^v=j} &= \begin{pmatrix} f_{X^*|W^v=j}(0) & 0 \\ 0 & f_{X^*|W^v=j}(1) \end{pmatrix}, \end{aligned}$$

where the superscript T stands for the transpose of a matrix. Similarly define the matrix representations $L_{X_a,W_a^u|W_a^v=j}$, $L_{X_a|X_a^*}$, $L_{W_a^u|X_a^*,W_a^v=j}$, and $L_{X_a^*|W_a^v=j}$ of the corresponding densities $f_{X_a,W_a^u|W_a^v=j}$, $f_{X_a|X_a^*}$, $f_{W_a^u|X_a^*,W_a^v=j}$ and $f_{X_a^*|W_a^v=j}$ in the auxiliary sample. Note that equation (40) implies for $j = 0, 1$,

$$\begin{aligned} & L_{X|X^*}L_{X^*|W^v=j}L_{W^u|X^*,W^v=j} \\ = \; & L_{X|X^*}\begin{pmatrix} f_{X^*|W^v=j}(0) & 0 \\ 0 & f_{X^*|W^v=j}(1) \end{pmatrix}\begin{pmatrix} f_{W^u|X^*,W^v=j}(0|0) & f_{W^u|X^*,W^v=j}(0|1) \\ f_{W^u|X^*,W^v=j}(1|0) & f_{W^u|X^*,W^v=j}(1|1) \end{pmatrix}^T \\ = \; & L_{X|X^*}\begin{pmatrix} f_{W^u,X^*|W^v=j}(0,0) & f_{W^u,X^*|W^v=j}(1,0) \\ f_{W^u,X^*|W^v=j}(0,1) & f_{W^u,X^*|W^v=j}(1,1) \end{pmatrix} \\ = \; & \begin{pmatrix} f_{X|X^*}(0|0) & f_{X|X^*}(0|1) \\ f_{X|X^*}(1|0) & f_{X|X^*}(1|1) \end{pmatrix}\begin{pmatrix} f_{W^u,X^*|W^v=j}(0,0) & f_{W^u,X^*|W^v=j}(1,0) \\ f_{W^u,X^*|W^v=j}(0,1) & f_{W^u,X^*|W^v=j}(1,1) \end{pmatrix} \\ = \; & \begin{pmatrix} f_{X,W^u|W^v=j}(0,0) & f_{X,W^u|W^v=j}(0,1) \\ f_{X,W^u|W^v=j}(1,0) & f_{X,W^u|W^v=j}(1,1) \end{pmatrix} \\ = \; & L_{X,W^u|W^v=j} \ , \end{aligned}$$

that is

$$L_{X,W^u|W^v=j} = L_{X|X^*}L_{X^*|W^v=j}L_{W^u|X^*,W^v=j}. \tag{42}$$

Similarly, equation (41) implies that

$$L_{X_a,W_a^u|W_a^v=j} = L_{X_a|X_a^*}L_{X_a^*|W_a^v=j}L_{W^u|X^*,W^v=j}. \tag{43}$$

The authors assume that *the observable matrices $L_{X_a,W_a^u|W_a^v=j}$ and $L_{X,W^u|W^v=j}$ are invertible, that the diagonal matrices $L_{X^*|W^v=j}$ and $L_{X_a^*|W_a^v=j}$ are invertible, and that $L_{X_a|X_a^*}$*

is invertible. Then equations (42) and (43) imply that $L_{X|X^*}$ and $L_{W^u|X^*,W^v=j}$ are invertible, and one can then eliminate $L_{W^u|X^*,W^v=j}$, to have for $j=0,1$

$$L_{X_a,W_a^u|W_a^v=j}L_{X,W^u|W^v=j}^{-1} = L_{X_a|X_a^*}L_{X_a^*|W_a^v=j}L_{X^*|W^v=j}^{-1}L_{X|X^*}^{-1}.$$

Since this equation holds for $j=0,1$, one may then eliminate $L_{X|X^*}$, to have

$$\begin{aligned} L_{X_a,X_a} &\equiv \left(L_{X_a,W_a^u|W_a^v=1}L_{X,W^u|W^v=1}^{-1}\right)\left(L_{X_a,W_a^u|W_a^v=0}L_{X,W^u|W^v=0}^{-1}\right)^{-1} \\ &= L_{X_a|X_a^*}\left(L_{X_a^*|W_a^v=1}L_{X^*|W^v=1}^{-1}L_{X^*|W^v=0}L_{X_a^*|W_a^v=0}^{-1}\right)L_{X_a|X_a^*}^{-1} \\ &\equiv \begin{pmatrix} f_{X_a|X_a^*}(0|0) & f_{X_a|X_a^*}(0|1) \\ f_{X_a|X_a^*}(1|0) & f_{X_a|X_a^*}(1|1) \end{pmatrix}\begin{pmatrix} k_{X_a^*}(0) & 0 \\ 0 & k_{X_a^*}(1) \end{pmatrix}\times \qquad (44) \\ &\quad \times\begin{pmatrix} f_{X_a|X_a^*}(0|0) & f_{X_a|X_a^*}(0|1) \\ f_{X_a|X_a^*}(1|0) & f_{X_a|X_a^*}(1|1) \end{pmatrix}^{-1}. \end{aligned}$$

with

$$k_{X_a^*}(x^*) = \frac{f_{X_a^*|W_a^v=1}(x^*)\,f_{X^*|W^v=0}(x^*)}{f_{X^*|W^v=1}(x^*)\,f_{X_a^*|W_a^v=0}(x^*)}.$$

Notice that the matrix $\left(L_{X_a^*|W_a^v=1}L_{X^*|W^v=1}^{-1}L_{X^*|W^v=0}L_{X_a^*|W_a^v=0}^{-1}\right)$ is diagonal because $L_{X^*|W^v=j}$ and $L_{X_a^*|W_a^v=j}$ are diagonal matrices. The equation (44) provides an eigenvalue-eigenvector decomposition of an observed matrix L_{X_a,X_a} on the left-hand side.

The authors assume that $k_{X_a^*}(0) \neq k_{X_a^*}(1)$*; i.e., the eigenvalues are distinctive.* This assumption requires that the distributions of the latent education level of males or females in the primary sample are different from those in the auxiliary sample, and that the distribution of the latent education level of males is different from that of females in one of the two samples. Notice that each eigenvector is a column in $L_{X_a|X_a^*}$, which is a conditional density. That means each eigenvector is automatically normalized. Therefore, for an observed L_{X_a,X_a}, one may have an eigenvalue-eigenvector decomposition as follows:

$$\begin{aligned} L_{X_a,X_a} &= \begin{pmatrix} f_{X_a|X_a^*}(0|x_1^*) & f_{X_a|X_a^*}(0|x_2^*) \\ f_{X_a|X_a^*}(1|x_1^*) & f_{X_a|X_a^*}(1|x_2^*) \end{pmatrix}\begin{pmatrix} k_{X_a^*}(x_1^*) & 0 \\ 0 & k_{X_a^*}(x_2^*) \end{pmatrix}\times \qquad (45) \\ &\quad \times\begin{pmatrix} f_{X_a|X_a^*}(0|x_1^*) & f_{X_a|X_a^*}(0|x_2^*) \\ f_{X_a|X_a^*}(1|x_1^*) & f_{X_a|X_a^*}(1|x_2^*) \end{pmatrix}^{-1}. \end{aligned}$$

The value of each entry on the right-hand side of equation (45) can be directly computed from the observed matrix L_{X_a,X_a}. The only ambiguity left in equation (45) is the value of the indices x_1^* and x_2^*, or the indexing of the eigenvalues and eigenvectors. In other words, the identification of $f_{X_a|X_a^*}$ boils down to finding a 1-to-1 mapping between the two sets of indices of the eigenvalues and eigenvectors: $\{x_1^*, x_2^*\} \Longleftrightarrow \{0, 1\}$.

Next, the authors make a normalization assumption that *people with (or without) college education in the auxiliary sample are more likely to report that they have (or do not have) college education; i.e., $f_{X_a|X_a^*}(x^*|x^*) > 0.5$ for $x^* = 0, 1$.* (This assumption also implies the invertibility of $L_{X_a|X_a^*}$.) Since the values of $f_{X_a|X_a^*}(0|x_1^*)$ and $f_{X_a|X_a^*}(1|x_1^*)$ are known in equation (45), this assumption pins down the index x_1^* as follows:

$$x_1^* = \begin{cases} 0 & \text{if } f_{X_a|X_a^*}(0|x_1^*) > 0.5 \\ 1 & \text{if } f_{X_a|X_a^*}(1|x_1^*) > 0.5 \end{cases} .$$

The value of x_2^* may be found in the same way. In summary, the authors have identified $L_{X_a|X_a^*}$, i.e., $f_{X_a|X_a^*}$, from the decomposition of the observed matrix L_{X_a,X_a}.

The authors then identify $L_{W^u|X^*,W^v=j}$ or $f_{W^u|X^*,W^v=j}$ from equation (43) as follows:

$$L_{X_a^*|W_a^v=j} L_{W^u|X^*,W^v=j} = L_{X_a|X_a^*}^{-1} L_{X_a,W_a^u|W_a^v=j},$$

in which two matrices $L_{X_a^*|W_a^v=j}$ and $L_{W^u|X^*,W^v=j}$ can be identified through their product on the left-hand side. Moreover, the density $f_{X|X^*}$ or the matrix $L_{X|X^*}$ is identified from equation (42) as follows:

$$L_{X|X^*} L_{X^*|W^v=j} = L_{X,W^u|W^v=j} L_{W^u|X^*,W^v=j}^{-1},$$

in which one may identify two matrices $L_{X|X^*}$ and $L_{X^*|W^v=j}$ from their product on the left-hand side. Finally, the density of interest $f_{X^*,W^u,W^v,Y}$ is identified from equation (39).

4.2 Models of continuous variables with nonclassical errors

Very recently there are a few papers address the identification and estimation of nonlinear EIV models in which continuous regressors are measured with arbitrarily nonclassical errors. For example, Hu and Schennach (2006) extend the method of Hu (2006) for misclassification of discrete regressors via IV approach to nonlinear models with a continuous regressor measured with a nonclassical error. Chen and Hu (2006) provide identification and estimation of nonlinear models with a continuous regressor measured with a nonclassical error via the two sample approach. Since the identification results of these two papers are extensions of those described in previous subsection for discrete regressor cases, we shall not discuss them here.

In order to obtain consistent estimates of the parameters β in the moment conditions $E[m(Y^*;\beta)] = 0$, **Chen, Hong, and Tamer (2005)** and **Chen, Hong, and Tarozzi (2004)** make use of an auxiliary data set to recover the correlation between the measurement errors and the underlying true variables by estimating the conditional distribution of the measurement errors given the observed reported variables or proxy variables. In their model, the auxiliary data set is a subset of the primary data, indicated by a dummy variable $D = 0$, which contains both the reported variable Y and the validated true variable Y^*. Y^* is not observed in the rest of the primary data set ($D = 1$) which is not validated. They assume that the conditional distribution of the true variables given the reported variables can be recovered from the auxiliary data set:

Assumption 11 $Y^* \perp D \mid Y$.

Under this assumption, an application of the law of iterated expectations gives

$$E\left[m\left(Y^*;\beta\right)\right] = \int g\left(Y;\beta\right) f\left(Y\right) dY \quad \text{where} \quad g\left(Y;\beta\right) = E\left[m\left(Y^*;\beta\right)|Y, D=0\right].$$

This suggests a semiparametric GMM estimator for the parameter β. For each value of β in the parameter space, the conditional expectation function $g(Y;\beta)$ can be nonparametrically estimated using the auxiliary data set where $D = 0$.

Chen, Hong, and Tamer (2005) use the method of sieves to implement this nonparametric regression. Let n denote the size of the entire primary dataset and let n_a denote the size of the auxiliary data set where $D = 0$. Let $\{q_l(Y), l = 1, 2, ...\}$ denote a sequence of known basis functions that can approximate any square-measurable function of X arbitrarily well. Also let

$$q^{k(n_a)}(Y) = \left(q_1(Y), ..., q_{k(n_a)}(Y)\right)' \quad \text{and}$$
$$Q_a = \left(q^{k(n_a)}(Y_{a1}), ..., q^{k(n_a)}(Y_{an_a})\right)'$$

for some integer $k(n_a)$, with $k(n_a) \to \infty$ and $k(n_a)/n \to 0$ when $n \to \infty$. In the above Y_{aj} denotes the jth observation in the auxiliary sample. Then for each given β, the first step nonparametric estimation can be defined as,

$$\hat{g}\left(Y;\beta\right) = \sum_{j=1}^{n_a} m\left(Y_{aj}^*;\beta\right) q^{k(n_a)}\left(Y_{aj}\right)\left(Q_a' Q_a\right)^{-1} q^{k(n_a)}\left(Y\right).$$

A GMM estimator for β_0 can then be defined using a positive definite weighting matrix $\hat{W}$ as

$$\hat{\beta} = \arg\min_{\beta \in B} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}\left(Y_i;\beta\right)\right)' \hat{W} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}\left(Y_i;\beta\right)\right).$$

Chen, Hong, and Tarozzi (2004) show that a proper choice of $\hat{W}$ achieves the semiparametric efficiency bound for the estimation of β. They called this estimator the *conditional expectation projection GMM* estimator.

Assumption 11 allows the auxiliary data set to be collected using a **stratified sampling** design where a *nonrandom response based subsample* of the primary data is validated. In a typical example of this stratified sampling design, we first oversample a certain subpopulation of the mismeasured variables Y, and then validate the true variables Y^* corresponding to this nonrandom stratified subsample of Y. It is very natural and sensible to oversample a subpopulation of the primary data set where more severe measurement error is suspected to be present. Assumption 11 is valid as long as in this sampling procedure of the auxiliary data set, the sampling scheme of Y in the auxiliary data is based only on the information available in the distribution of the primary data set $\{Y\}$. For example, one can choose a subset of the primary data set $\{Y\}$ and validate the corresponding $\{Y^*\}$, in which case the Y's in the auxiliary data set are a subset of the primary data Y. The stratified sampling procedure can be illustrated as follows. Let U_{pi} be i.i.d $U(0,1)$ random variables independent of both Y_{pi} and Y_{pi}^*, and let $T\left(Y_{pi}\right) \in (0,1)$ be a measurable function of the primary data. The stratified sample is obtained by validating every observation for which $U_{pi} < T\left(Y_{pi}\right)$. In other words, $T\left(Y_{pi}\right)$ specifies the probability of validating an observation after Y_{pi} is observed.

A special case of assumption 11 is when the auxiliary data is generated from the same population as the primary data, where a full independence assumption is satisfied:

Assumption 12 $Y, Y^* \perp D$.

This case is often referred to as a (true) **validation sample**. Semiparametric estimators that make use of a validation sample include Carroll and Wand (1991), Sepanski and Carroll (1993), Lee and Sepanski (1995) and the recent work of Devereux and Tripathi (2005). Interestingly, in the case of a validation sample, Lee and Sepanski (1995) suggested that the nonparametric estimation of the conditional expectation function $g\left(Y;\beta\right)$ can be replaced by a finite dimensional linear projection $h\left(Y;\beta\right)$ into a fixed set of functions of Y. In other words, instead of requiring that $k(n_a) \to \infty$ and $k(n_a)/n \to 0$, we can hold $k\left(n_a\right)$ to be a fixed constant in the above least square regression for $\hat{g}\left(Y;\beta\right)$. Lee and Sepanski (1995) show that this still produces a consistent and asymptotically normal estimator for β as long as the auxiliary sample is also a validation sample that satisfies assumption 12. However, if the auxiliary sample satisfies only assumption 11 but not assumption 12, then it is necessary to require $k(n_a) \to \infty$ to obtain consistency. Furthermore, even in the case of a validation sample, requiring $k(n_a) \to \infty$ typically leads to a more efficient estimator for β than a constant $k(n_a)$.

An alternative consistent estimator that is valid under assumption 11 is based on the inverse probability weighting principle which provides an equivalent representation of the moment condition $Em\left(y^*;\beta\right)$. Define $p\left(Y\right) = p\left(D=1|Y\right)$,

$$Em\left(y^*;\beta\right) = E\left[m(Y^*;\beta_0)\frac{1-p}{1-p(Y)}\ \middle|\ D=0\right].$$

To see this, note that,

$$\begin{aligned}
&E\left[m(Y^*;\beta_0)\frac{1-p}{1-p(Y)}\ \middle|\ D=0\right] \\
=&\int m(Y^*;\beta_0)\frac{1-p}{1-p(Y)}\frac{f\left(Y\right)\left(1-p\left(Y\right)\right)f\left(Y^*|Y,D=0\right)}{1-p}dY^*dY \\
=&\int m(Y^*;\beta_0)\,f\left(Y^*|Y\right)f\left(Y\right)\,dY^*dY = E\;m\left(y^*;\beta\right),
\end{aligned}$$

where the third equality follows from assumption 11 that $f\left(Y^*|Y,D=0\right) = f\left(Y^*|Y\right)$.

This equivalent reformulation of the moment condition $E\;m\left(Y^*;\beta\right)$ suggests a two-step *inverse probability weighting GMM* estimation procedure. In the first step, one typically obtains a parametric or nonparametric estimate of the so-called propensity score $\widehat{p}\,(Y)$ using for example a logistic binary choice model with a flexible functional form. In the second step, a sample analog of the re-weighted moment conditions is computed using the auxiliary data set:

$$\widehat{g}\,(\beta) = \frac{1}{n_a} \sum_{j=1}^{n_a} m\left(Y_j^*; \beta\right) \frac{1}{1 - \widehat{p}\,(Y_j)}.$$

This is then used to form a quadratic norm to provide a GMM estimator:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, \widehat{g}\,(\beta)' \, W_n \widehat{g}\,(\beta)\,.$$

The authors then apply their estimator to study the returns to schooling as the influence of the number years of schooling on the individual earning. The data used for estimation are taken from the Current Population Survey matched with employer-reported (or from the social security records) social security earnings. As the social security data provide more accurate information about individual incomes but cannot be matched to all individuals in the sample, the authors use the social security records to form a validation sample. The standard Mincer model is used to study the dependence of the logarithm of individual income on education, experience, experience squared, and race. The objective function that defines their estimator is built from the least absolute deviation estimator of Powell (1984) (allowing them to "filter out" censoring caused by the top coding of the social security data), which is projected to the set of observed variables: mismeasured income, education, experience, and race. The authors use sieves to make such projection, by representing the data density by the sieve expansion and approximating integration by summation. Then they obtain the estimates from the conventional LAD estimation for the primary and auxiliary samples, and the estimates obtained using the method suggested in the paper. They found a significant discrepancy (almost 1%) between the return to education obtained from the primary sample and the estimates from the suggested method.

Interestingly, an analog of the conditional independence assumption 11 is also rooted in the program evaluation literature and is typically referred to as the assumption of unconfoundedness, or selection based on observable. Semi-parametric efficiency results for the mean treatment effect parameters to nonlinear GMM models have been developed by, among other, Robins, Mark, and Newey (1992), Hahn (1998) and Hirano, Imbens, and Ridder (2003). Many of the results presented here generalize these results for the mean treatment effect parameters to nonlinear GMM models.

An example of GMM-based estimation procedure which achieves the semiparametric efficiency bound can be found in Chen, Hong, and Tarozzi (2004). Given Assumption 11 the authors provide a methodology for parameter estimation in the semiparametric framework and describe the structure of the asymptotic distribution of the obtained estimator. Let us consider this paper in more detail. Under Assumption 11, the authors follow the framework of Newey (1990) to show that the efficiency bound for estimating β is given by $\left(\mathcal{J}_\beta' \Omega_\beta^{-1} \mathcal{J}_\beta\right)^{-1}$, where for $p\left(Y\right) = p\left(D=1|Y\right)$:

$$\mathcal{J}_\beta = \frac{\partial}{\partial \beta} E\left[m\left(Y^*;\beta\right)\right] \quad \text{and} \quad \Omega_\beta = E\left[\frac{1}{1-p\left(Y\right)} V\left[m\left(Y^*;\beta\right) \mid Y\right] + \mathcal{E}\left(Y;\beta\right)\mathcal{E}\left(Y;\beta\right)'\right].$$

We can demonstrate this result in three steps. First we characterize the properties of the tangent space under assumption 11. Next we write the parameter of interest in its differential form and therefore find a linear influence function d. Finally we conjecture and verify the projection of d onto the tangent space and the variance of this projection gives rise to the efficiency bound. We first go through these three steps under the assumption that the moment conditions exactly identify β. Finally, the results are extended to overidentified moment conditions by considering their optimal linear combinations.

First we assume that the moment conditions exactly identify β.

Step 1. Consider a parametric path θ of the joint distribution of Y, Y^* and D. Define $p_\theta\left(y\right) = P_\theta\left(D=1|y\right)$. Under assumption 1, the joint density function for Y^*, D and Y can be factorized into

$$f_\theta\left(y^*, y, d\right) = f_\theta\left(y\right) p_\theta\left(y\right)^d \left[1-p_\theta\left(y\right)\right]^{1-d} f_\theta\left(y^* \mid y\right)^{1-d}. \tag{46}$$

The resulting score function is then given by

$$S_\theta\left(d, y^*, y\right) = \left(1-d\right) s_\theta\left(y^* \mid y\right) + \frac{d - p_\theta\left(y\right)}{p_\theta\left(y\right)\left(1-p_\theta\left(y\right)\right)} \dot{p}_\theta\left(y\right) + t_\theta\left(x\right),$$

where

$$s_\theta\left(y^* \mid y\right) = \frac{\partial}{\partial \theta} \log f_\theta\left(y^* \mid y\right), \quad \dot{p}_\theta\left(y\right) = \frac{\partial}{\partial \theta} p_\theta\left(y\right), \quad t_\theta\left(y\right) = \frac{\partial}{\partial \theta} \log f_\theta\left(y\right) ...$$

The tangent space of this model is therefore given by:

$$\mathcal{T} = \left\{ (1-d)\, s_\theta\left(y^* \mid y\right) + a\left(y\right)\left(d - p_\theta\left(y\right)\right) + t_\theta\left(y\right) \right\}, \tag{47}$$

where $\int s_\theta\left(y^* \mid y\right) f_\theta\left(y^* \mid y\right) dy = 0$, $\int t_\theta\left(y\right) f_\theta\left(y\right) dy = 0$, and $a\left(y\right)$ is any square integrable function.

Step 2. As in the method of moment model in Newey (1990), the differential form of the parameter β can be written as

$$\begin{aligned} \frac{\partial \beta\left(\theta\right)}{\partial \theta} &= -\left(\mathcal{J}_\beta\right)^{-1} E\left[m\left(Y^*;\beta\right) \frac{\partial \log f_\theta\left(Y^*, Y\right)}{\partial \theta'} \right] \\ &= -\left(\mathcal{J}_\beta\right)^{-1} \left\{ E\left[m\left(Y^*;\beta\right) \left(s_\theta\left(Y^* \mid Y\right)' + t_\theta\left(Y\right)' \right) \right] \right\} \\ &= -\left(\mathcal{J}_\beta\right)^{-1} \left\{ E\left[m\left(Y^*;\beta\right) s_\theta\left(Y^* \mid Y\right)' \right] + E\left[\mathcal{E}\left(Y\right) t_\theta\left(Y\right)' \right] \right\}. \end{aligned} \tag{48}$$

Therefore $d = -\mathcal{J}_\beta^{-1} m\left(Y^*;\beta\right)$. Since J_β is only a constant matrix of nonsingular transformation. The projection of d onto the tangent space will be $-\mathcal{J}_\beta$ multiplied by the projection of $m\left(Y^*;\beta\right)$ onto the tangent space. Therefore we only need to consider the projection of $m\left(Y^*;\beta\right)$ onto the tangent space.

Step 3. We conjecture that this projection takes the form of

$$\tau\left(Y^*, Y, D\right) = \frac{1-D}{1-p(Y)} \left[m\left(Y^*;\beta\right) - \mathcal{E}\left(Y\right) \right] + \mathcal{E}\left(Y\right).$$

To verify that this is the efficient influence function we need to check that $\tau\left(Y^*, Y, D\right)$ lies in the tangent space and that

$$E\left[\left(m\left(Y^*;\beta\right) - \tau\left(Y^*, Y, D\right) \right) s_\theta\left(Y^*, Y\right) \right] = 0,$$

or that

$$E\left[m\left(Y^*;\beta\right) s_\theta\left(Y^*, X\right) \right] = E\left[\tau\left(Y^*, Y, D\right) s_\theta\left(Y^*, Y\right) \right]. \tag{49}$$

To see that $\tau\left(Y^*, Y, D\right)$ lies in the tangent space, note that the first term in $\tau\left(Y^*, Y, D\right)$ has mean zero conditional on X, and corresponds to the first term of $(1-d)\, s_\theta\left(y^* | y\right)$ in the tangent space. The second term in $\tau\left(Y^*, Y, D\right)$, $\mathcal{E}\left(y\right)$, has unconditional mean zero and obviously corresponds to the $t_\theta\left(y\right)$ in the tangent space.

To verify (49), one can make use of the representation of $E\left[m\left(Y^*;\beta\right)s_\theta\left(Y^*,Y\right)\right]$ in (48), by verifying the two terms in $\tau\left(Y^*,Y,D\right)$ separately. The second term is obvious and tautological. The first part,

$$E\left[\frac{1-D}{1-p(Y)}\left[m\left(Y^*;\beta\right)-\mathcal{E}\left(Y\right)\right]s_\theta\left(Y^*,Y\right)\right]=E\left[m\left(Y^*;\beta\right)s_\theta\left(Y^*,Y\right)\right],$$

follows from the conditional independence assumption 11 and the score function property $E\left[s_\theta\left(Y^*,Y\right)|Y\right]=0$. Therefore we have verified that $\tau\left(Y^*,Y,D\right)$ is the efficient projection and that the efficiency bound is given by

$$\begin{aligned} V &= \left(\mathcal{J}_\beta\right)^{-1}E\left[\tau\left(Y^*,Y,D\right)\tau\left(Y^*,Y,D\right)'\right]\left(\mathcal{J}_\beta\right)'^{-1} \\ &= \left(\mathcal{J}_\beta\right)^{-1}E\left[\frac{1}{1-p(Y)}Var\left(m\left(Y^*;\beta\right)\mid Y\right)+\mathcal{E}\left(Y\right)\mathcal{E}\left(Y\right)'\right]\left(\mathcal{J}_\beta\right)'^{-1}. \end{aligned}$$

Finally consider the extensions of these results to the overidentified case. When $d_m>d_\beta$, the moment condition is equivalent to the requirement that for any matrix $\mathcal{A}$ of dimension $d_\beta\times d_m$ the following exactly identified system of moment conditions holds

$$\mathcal{A}E\left[m\left(Y^*;\beta\right)\right]=0.$$

Differentiating under the integral again, we have

$$\frac{\partial\beta\left(\theta\right)}{\partial\theta}=-\left(\mathcal{A}E\left[\frac{\partial m\left(Y^*;\beta\right)}{\partial\beta}\right]\right)^{-1}E\left[\mathcal{A}m\left(Y^*;\beta\right)\frac{\partial\log f_\theta\left(Y^*,Y\mid D=1\right)}{\partial\theta'}\right].$$

Therefore, any regular estimator for β will be asymptotically linear with influence function of the form

$$-\left(\mathcal{A}E\left[\frac{\partial m\left(Y^*;\beta\right)}{\partial\beta}\right]\right)^{-1}\mathcal{A}m\left(Y^*;\beta\right).$$

For a given matrix $\mathcal{A}$, the projection of the above influence function onto the tangent set follows from the previous calculations, and is given by

$$-\left[\mathcal{A}\mathcal{J}_\beta\right]^{-1}\mathcal{A}\tau\left(y,x,d\right).$$

The asymptotic variance corresponding to this efficient influence function for fixed $\mathcal{A}$ is therefore

$$\left[\mathcal{A}\mathcal{J}_\beta\right]^{-1}\mathcal{A}\Omega\mathcal{A}'\left[\mathcal{J}_\beta\mathcal{A}'\right]^{-1} \tag{50}$$

where

$$\Omega = E\left[\tau\left(Y^*, Y, D\right)\tau\left(Y^*, Y, D\right)'\right]$$

as calculated above. Therefore, the efficient influence function is obtained when $\mathcal{A}$ is chosen to minimize this efficient variance. It is easy to show that the optimal choice of $\mathcal{A}$ is equal to $\mathcal{J}_\beta'\Omega^{-1}$, so that the asymptotic variance becomes

$$V = \left(\mathcal{J}_\beta'\Omega^{-1}\mathcal{J}_\beta\right)^{-1}.$$

Different estimation methods can be used to achieve this semiparametric efficiency bound. In particular, Chen, Hong, and Tarozzi (2004) show that both a semiparametric conditional expectation projection estimator and a semiparametric propensity score estimator based on a sieve nonparametric first stage regression achieve this efficiency bound.

Other recent papers that develop estimation methods using combined samples include Linton and Whang (2002), Devereux and Tripathi (2005), Ichimura and Martinez-Sanchis (2005) and Hu and Ridder (2006).

5 Contaminated and Corrupted Data

Instead of assuming an additive measurement error term, another model of data mismeasurement is to assume that every observation can be contaminated or corrupted with a certain probability. A literature of robust estimation, in the spirit of Horowitz and Manski (1995), aims at providing estimators that are robust to the presence of such errors in the data. Without strong identifying assumption, population parameters are not necessarily point identified. But if one uses robust estimation methods, bound information can potentially be obtained for these parameters.

Horowitz and Manski (1995) provide consistent and sharp bounds on the latent distribution function when data is subject to contamination under the only assumption that an upper bound can be put on the probability of the data error. They consider the problem of estimation of the marginal distribution of the random variable y_0, while the actually observed variable is y, which is a proxy for y_0 and is contaminated by a measurement error.

In their model,

$$y \equiv y_0(1-d) + y_1 d,$$

where y_1 is the erroneous response, y_0 is the true response and $d \in \{0, 1\}$. On the one hand, if $d = 0$, then the observation of the response variable $y = y_0$ is free of error. On the other hand, if $d = 1$, then $y = y_1$ is a contaminated observation. It is assumed that the variables y_0 and y_1 have a common support Y.

Denote $Q \equiv Q(y)$ the distribution of y and use $P_i \equiv P_i(y_i)$ to denote the marginal distribution of y_i $(i = 0, 1)$. In addition $P_{ij} \equiv P(y_i \mid d = j)$ is the distribution of the variables y_i $(i = 0, 1)$ conditional on d $(j = 0, 1)$. Then $p \equiv P(d = 1)$ is the marginal probability of the data contamination.

We are interested in a parameter $\tau(P_0)$ where τ maps the space of probability distributions Ψ on (Y, Ω) into $\mathbb{R}$. The data does not reveal complete information about this distribution. The distribution which can be observed from the data is:

$$Q = (1-p)\, P_{00} + pP_{11}.$$

We are interested, though, in estimating the the marginal distribution of the true response:

$$P_0 = (1-p)P_{00} + pP_{01}.$$

In general, if the prior information about the misclassification probability p is not available, then the observable distribution Q does not impose any restrictions on the unobservable distribution.

A common assumption easing the identification problem is that the occurrence of data errors and the sample realization are independent: $P_0 = P_{01}$. This assumption in general allows one to provide tighter bounds on the probability of the true response P_0. In addition to this, one can assume that the probability of the data error, p, is bounded from above by some constant $\lambda < 1$. The authors mention that in general, we can have a consistent estimate of λ and conduct the analysis assuming exact prior information about this parameter.

For a given λ, it is possible to obtain a set of probability distributions containing the distribution of the true response. Suppose first that for a given p we construct bounds for the marginal distribution of the true response P_0 and the distribution of the true response given that the response is observed with error P_{01}. In this case the distribution of the true response given that we are observing the response without an error:

$$P_{00} \in \Psi_{00}(p) \equiv \Psi \cap \{(Q - p\psi_{11})/(1-p) \;:\; \psi_{11} \in \Psi\}\,.$$

This defines the set of distributions $\Psi_{00}(p)$ where we should expect to find the true response given the error-free observation. Based on this set we can confine the marginal distribution of the true response given that we observe the true response with probability p so that:

$$\begin{aligned} P_0 \in \Psi_0(p) \equiv & \Psi \cap \{(1-p)\,\psi_{00} + p\psi_{01} \;:\; (\psi_{00}, \psi_{01}) \in \Psi_{00}\,(p) \times \Psi\} \\ = & \Psi \cap \{Q - p\psi_{11} + p\psi_{01} \;:\; (\psi_{11}, \psi_{01}) \in \Psi \times \Psi\}\,. \end{aligned}$$

The authors show that the set $\Psi_{00}(p)$ is inside the set of distributions $\Psi_0(p)$.

Moreover, if the argument p is increasing, we obtain a monotone sequence of sets, such that for $\delta > 0$: $\Psi_0\,(p) \subset \Psi_0\,(p+\delta)$ and $\Psi_{00}\,(p) \subset \Psi_{00}\,(p+\delta)$. Specifically, the fact that $p \leq \lambda < 1$ implies that $P_{00} \in \Psi_{00}\,(\lambda)$ and $P_0 \in \Psi_0\,(\lambda)$. In case when the errors are independent from the binary outcomes, then the sets $\Psi_0(\lambda)$ and $\Psi_{00}(\lambda)$ coincide, which means that P_0 belongs to a smaller set $\Psi_{00}(\lambda)$.

Given the identification results for the marginal distribution of the binary outcome, the authors extend the results to the case of a general estimator $\tau(\cdot)$, considered as a real-valued function defined on the family of distributions Ψ. Denote the set of values of the estimator on the set of probability distributions Ψ by $T = (T_L, T_U)$. The set of parameter values as a function of the set of distributions of the true response yt_0 can be written as:

$$\tau(P_{00}) \in T_{00}(\lambda) \equiv \{\tau(\psi) \;:\; \psi \in \Psi_{00}(\lambda)\}$$

and

$$\tau(P_0) \in T_0(\lambda) \equiv \{\tau(\psi) \;:\; \psi \in \Psi_0(\lambda)\}\,.$$

The shape of these sets depends on λ. Let $T_{11L}(\lambda)$ and $T_{11U}(\lambda)$ denote the lower and upper bound of $T_{11}(\lambda)$. Let $T_{1L}(\lambda)$ and $T_{1U}(\lambda)$ denote the lower and upper bounds of $T_1(\lambda)$.

We can find the maximum probability of observing the erroneous outcome λ such that the estimator $\tau(\cdot)$ is not shifted to the boundary of its range as:

$$\lambda_{00} \equiv \sup\left\{\lambda : T_L < T_{00L}(\lambda) \leq T_{00U}(\lambda) < T_U\right\},$$

and

$$\lambda_0 \equiv \sup\left\{\lambda : T_L < T_{0L}(\lambda) \leq T_{0U}(\lambda) < T_U\right\}.$$

The authors call these values λ_0 and λ_{00} the "identification breakdown" points of $\tau(\cdot)$. This notion suggests that at these values of the probability of data corruption the information from the data does not allow us to make the *a priori* information more precise.

The bounds on the probability distributions P_{00} and P_0 of the true response variable can be transformed to statements about the probability of some outcome configuration A (in the σ-algebra on Ω). The locus of the identified sets is described by a simple intersection of the segments:

$$P_{00}(A) \in \Psi_{00}\left(A;\lambda\right) \equiv [0,\,1] \cap \left[\frac{Q(A)-\lambda}{1-\lambda}\,,\;\frac{Q(A)}{1-\lambda}\right]$$

and

$$P_0(A) \in \Psi_0\left(A;\lambda\right) \equiv [0,\,1] \cap \left[Q(A)-\lambda\;,\;Q(A)+\lambda\right].$$

These expressions imply that the probability of observing the outcome A when the data are contaminated by error is inside the set $\Psi_{00}\left(A;\lambda\right)$ which is equal to the entire segment $[0,\,1]$ if $1-\lambda \leq Q(A) \leq \lambda$. When λ is small enough compared to $Q\left(A\right)$, a possible set of values of the probability of realization of y is smaller that the entire interval.

A more concrete discussion of the bounds can be made in the case when we consider a continuous set of values of the random outcome y. Specifically, suppose that Y is the extended real line and Ω consists of Lebesgue measurable sets. In this case the bounds provided for the family of probability distributions for the correct outcome can be used to build the bounds for the quantiles of the distribution of y. The α-quantiles of P_{00} and P_0 are respectively $q_{00}(\alpha) = \inf\left\{t\,:\,P_{00}\left[-\infty,t\right] \geq \alpha\right\}$ and $q_0(\alpha) = \inf\left\{t\,:\,P_0\left[-\infty,t\right] \geq \alpha\right\}$.

To state results for the bounds of the quantiles of the distribution of the true outcome, we first introduce the function:

$$r(\gamma) = \begin{cases} \gamma - \text{ quantile of } Q \text{ if } 0 < \gamma \leq 1, \\ -\infty \text{ if } \gamma \leq 0, \\ +\infty \text{ if } \gamma > 1. \end{cases}$$

Specifically, the α-quantiles will be inside the following segments:

$$q_{00}(\alpha) \in [r(\alpha(1-\lambda)), r(\alpha(1-\lambda)+\lambda)]$$

and

$$q_0(\alpha) \in [r(\alpha-\lambda), r(\alpha+\lambda)].$$

Note that as the bound for probability λ increases, the bounds for the quantiles become wider. It approaches the entire support of the observed distribution Q when the bound λ exceeds the identification breakdown point.

The authors also consider the case when the estimator τ respects the stochastic dominance of the distributions. This means that if the distribution F fist-order stochastically dominates the distribution G then $\tau(F) \geq \tau(G)$. In this case using the definition of the quantile function $r(\cdot)$ introduced above, we can introduce the auxiliary functions:

$$L_\lambda[-\infty, t] = \begin{cases} Q[-\infty, t]/(1-\lambda) \quad \text{if } t < r(1-\lambda), \\ 1 \quad \text{if } t \geq r(1-\lambda), \end{cases}$$

$$U_\lambda[-\infty, t] = \begin{cases} 0 \quad \text{if } t < r(\lambda), \\ (Q[-\infty, t] - \lambda)/(1-\lambda) \quad \text{if } t \geq r(\lambda), \end{cases}$$

In this case we can express the bounds for the parameter $\tau(\cdot)$ as a function of the true latent distribution as:

$$\tau(P_{00}) \in [\tau(L_\lambda), \tau(U_\lambda)].$$

Using the previously introduced notion of a δ-function we can define the bounds for the values of the estimator for the marginal distribution of the true outcome as:

$$\tau(P_0) \in [\tau\{(1-\lambda)L_\lambda + \lambda\delta_{-\infty}\}, \tau\{(1-\lambda)U_\lambda + \lambda\delta_{+\infty}\}].$$

One of the implications of this result for the general case of the interval for the estimators respecting the stochastic dominance is a practically relevant case when estimator takes the form $\tau(\psi) = \int g(y)\, d\psi$ for ψ- the distribution of the outcome and a function $g(\cdot)$ with a limit at positive infinity equal to K_1 and the limit at negative infinity equal to K_0. In this case the bounds for the values of the estimator will be determined by the integrals over the weighted distribution functions L_λ and U_λ. Specifically one can write that:

$$\tau\left(P_{00}\right) \in \left[\int g(y)\, dL_\lambda,\ \int g(y)\, dU_\lambda\right]$$

The expression for the values of the estimator for the entire marginal distribution of the true outcome employs the fact that δ-function works as a shifting operator for the kernel in the integral and thus:

$$\tau\left(P_0\right) \in \left[(1-\lambda)\int g(y) dL_\lambda + \lambda K_0,\ (1-\lambda)\int g(y) dU_\lambda + \lambda K_1\right].$$

One can see that the bounds for the estimator on the set of distributions of the true outcome given that the outcome is observed without an error do not depend on the asymptotic values of the function $g(\cdot)$. As a result, sharp bounds for the estimator of $\tau\left(P_{00}\right)$ can be obtained even if the kernel function $g(\cdot)$ is unbounded.

The final step of the authors allows them to provide local bounds for the estimator in case of smooth functionals $\tau(\cdot)$. First, note that the sets of distributions of the true response y_0 can be expressed in terms of the observable distribution Q:

$$\Psi_{00}\left(\lambda\right) = \left\{Q - \left[\lambda/(1-\lambda)\left(\psi - Q\right)\right],\ \text{where } \psi \in \Psi_{11}(\lambda)\right\}$$

and

$$\Psi_0\left(\lambda\right) = \left\{Q - \lambda\left(\psi - \omega\right),\ \text{where } \psi \in \Psi_{11}(\lambda),\ \omega \in \Psi\right\}.$$

Then a parameter τ with a general structure, we can define it as a functional $\tau\left(Q,\, \psi,\, \omega\right)$. For this functional one can define a directional derivative as:

$$\tau'\left(Q, \psi, \omega\right) = \lim_{\beta \downarrow 0} \frac{\tau\left[Q - \beta\left(\psi - \omega\right)\right] - \tau\left[Q\right]}{\beta},$$

which determines the sensitivity of the estimated parameter to the misclassification probability. If such derivative exists then it allows one to determine the bounds for the parameter estimate when the misclassification probability is infinitesimal. Specifically we can express the bounds for the estimator τ evaluated for the set of the distribution of the true response given that the observation does not contain error as:

$$\begin{aligned}\tau(Q) + \lambda \inf_{\psi \in \Psi_{11}(\lambda)} \tau' \left(Q, \psi, Q\right) + o\left(\lambda; Q\right) \leq & \tau\left(P_{00}\right) \\ & \leq \tau(Q) + \lambda \sup_{\psi \in \Psi_{11}(\lambda)} \tau' \left(Q, \psi, Q\right) + o\left(\lambda; Q\right).\end{aligned}$$

The bounds for the parameter evaluated at the marginal distribution of the true response are evaluated in the same way, but the upper and lower bounds are taken also over the set of possible distributions of the true response given that the erroneous response is observed. The above bounds rely on a strong assumption that the directional derivative exists and can be uniformly approximated in the selected subsets of response distributions. This limits the analysis to sufficiently smooth functionals.

A practically useful application of the infinitesimal bounds concerns the analysis of estimators which have local integral representation. Specifically, if we assume that we can locally express the directional derivative as $\tau'(Q,\, \psi,\, \omega) = \int f_Q(y) d\left(\psi - \omega\right)$, then the above bounds for the estimator in the integral representation can be expressed in terms of the upper and the lower bound of the kernel $f_Q(\cdot)$ on the support of y. Specifically, assuming that $\int f_Q(y)\, dQ = 0$ we can express the bounds locally as:

$$\tau\left(P_{00}\right) \in \left[\tau\left(Q\right) + \lambda \inf_{y \in Y} f_Q(y) + o\left(\lambda;\, Q\right),\ \tau\left(Q\right) + \lambda \sup_{y \in Y} f_Q(y) + o\left(\lambda;\, Q\right)\right].$$

Thus the identification of the estimator is determined by deviations of the values of the estimator for the true values of the variables from the estimator at the observed distribution driven by infinitesimal probability of data contamination.

Horowitz and Manski (1995) apply their methodology for evaluation of the bounds of the identified set of distribution to analyze the income distribution in the U.S. The authors use the data from the Current Population Survey and analyze the characteristics of the income distribution. Approximately 8% of the survey respondents provided incomplete data about their income and 4.5% of the respondents in the CPS were not interviewed. This allows the authors to provide a consistent estimate of the upper bound on the probability of erroneous response of 12.1%. The application of their method to this data set then allows them to obtain the bounds for the error-corrected quantiles of the income distribution.

Molinari (2005) uses a different approach to identifying the true outcome distribution from error-contaminated observations. The author uses the direct misclassification approach in which the true and the observed response are connected by a system of linear equations with the coefficients equal to the misclassification probabilities. The prior information is incorporated in the form of the functional restrictions on the element of the matrix of misclassification probabilities and it allows one to construct tight confidence intervals for various statistics of the true response.

Following the previous notation we use y_0 for the true outcome, y_1 for the erroneous outcome and y for the observed outcome. The binary variable d is equal to 0 when the true outcome is observed and one otherwise. The author assumes that the set of values of the true outcome Y is discrete and that the supports of both y_1 and y_0 are in Y. The author introduces the marginal distribution of the true outcome in the vector form $\mathbf{P}_0 = \left[P_0^j, \, j \in Y\right] \equiv \left[\Pr\left(y_0 = j\right), \, j \in Y\right]$ and the marginal distribution of the observed outcome in the vector form is $\mathbf{Q} = \left[Q^j, \, j \in Y\right] \equiv \left[\Pr\left(y = j\right), \, j \in Y\right]$.

In addition to this, the matrix of conditional probabilities for the observable response given the true response is defined as

$$\Pi^\star = \left(\pi_{ij}\right)_{i,\, j \in Y} \equiv \left(\Pr\left(y = i \mid y_0 = j\right)\right)_{i,\, j \in Y}.$$

The parameter of interest is a real-valued function on the space of probability distributions Ψ: $\tau\left[\mathbf{P}_0\right]$.

The marginal distribution of the observable outcome can be expressed in a matrix form in terms of the true outcome as

$$\mathbf{Q} = \Pi^\star \cdot \mathbf{P}_0$$

If the matrix of probabilities $\Pi^\star$ was known and had full rank, than we could retrieve the statistic $\tau\left(\mathbf{P}_0\right)$ from the probabilities of the observed outcome by inverting the matrix $\Pi^\star$.

This is usually not the case and the prior information comes in the form of the set of possible values of misclassification probabilities $H\left[\Pi^{\star}\right]$ for each element of this matrix.

The identification region for the distribution of the true outcome can be obtained from the observable distribution and the identification bounds as a set:

$$\Psi_0 = \left\{\psi \,:\, \mathbf{Q} = \Pi \cdot \psi,\; \Pi \in H\left[\Pi^{\star}\right]\right\}.$$

In the further discussion we will be using p_0 to denote a point of the identified set of distributions Ψ_0. The identification region for the statistic $\tau(\mathbf{P}_0)$ can be expressed as:

$$T_0 = \left\{\tau\left(\psi\right) \,:\, \psi \in \Psi_0\right\}.$$

If $H^P\left[\Pi^{\star}\right]$ is the set of matrices satisfying the probabilistic constraints and $H^E\left[\Pi^{\star}\right]$ is the set of matrices satisfying the constraints from validation studies, then the identification set for the matrix of misclassification probabilities is:

$$H\left[\Pi^{\star}\right] = H^P\left[\Pi^{\star}\right] \cap H^E\left[\Pi^{\star}\right]$$

Apparently, the geometry of the set $H\left[\Pi^{\star}\right]$ will be translated to the geometry of the set Ψ_0. Specifically, if the set of restrictions from validation studies is not connected, then the identification set of the statistic $\tau(\cdot)$ can also be unconnected.

The set of probabilistic restrictions implies that the matrix $\Pi^{\star}$ should be stochastic and that multiplication of the vector of probabilities $\mathbf{P}_0$ by this matrix should give a proper distribution $\mathbf{Q}$. In this case if Δ_n is an n-dimensional simplex and $conv\left(a_1, a_2, \ldots, a_n\right)$ is the convex hull of a collection of vectors $\{a_k\}_{k=1}^{n}$, then the set of probabilistic restrictions can be written as:

$$H^P\left[\Pi^{\star}\right] = \Big\{\Pi \,:\, \pi_j \in \Delta_{|Y|-1} \text{ and } \psi_0^j \geq 0, \forall j \in Y,\, \psi_0 \in \Psi_0, \text{ and } \mathbf{Q} \in conv\left(\pi_1, \ldots, \pi_{|Y|}\right)\Big\},$$

where π_j stands for the j-th column of matrix $\Pi^{\star}$. To further describe the properties of the set of the probabilistic restrictions the author uses the notion of star convexity. Star convexity of a certain set with respect to a specific point γ implies that for any point of this set and γ the line segment connecting these two points should lie inside the set.

The author proves that if $\widetilde{\Pi}$ is the matrix with all columns equal to the vector of observed probabilities $\mathbf{Q}$, then the set of probabilistic restrictions $H^P[\Pi^\star]$ is star convex with respect to $\widetilde{\Pi}$ but not star convex with respect to any of its other elements.

The author provides different examples of possible set of validation restrictions. One such restriction has been considered in the paper Horowitz and Manski (1995) outlined above, where one can impose the upper bound on the probabilities of erroneous outcome, and thus impose a lower bound restriction on the elements of the diagonal of the matrix $\Pi^\star$. The other example of such restrictions is when the variable y_0 tends to be over-reported which means that:

$$H^E[\Pi^\star] = \{\Pi \,:\, \pi_{ij} = 0,\ \forall i < j \in Y\}.$$

Despite the fact that the set of probabilistic restrictions is not convex, depending on the set of validation restrictions, the resulting identification set for the elements of $\Pi^\star$ can be convex, connected or disconnected.

The author then uses the technique for estimation of set - identified parameters to recover the elements of the true outcome distribution $\mathbf{P}_0$. The technique is based on treatment of restrictions for the elements in $H^E[\Pi^\star]$ as inequality and equality constraints given the probabilistic restrictions. Then the problem of verifying whether an element $\psi_0 \in \Psi_0$ satisfies the constraints reduces to the problem of looking for a feasible solution in a linear programming problem. The author proves that the identified region constructed in this way will be consistent in the supremum-norm sense for the true identified region.

Molinari (2005) uses the data from the Health and Retirement Study (HRS) to illustrate her identification and estimation methodology. The author studies the distribution of the types of pension plans in the population of the currently employed Americans for the period between the year 1992 and 1998. A significant inference problem is that in general the workers might be misinformed about the characteristics of their pension plans, and for this reason a substantial amount of error might be present in the survey data. The respondents have three pension plans available, and the author possesses an additional dataset which matches the individuals in the survey to the exact data provided by the Social Security Administration. This additional dataset is used to impose the restrictions on the matrix of misreporting probabilities. Then, assuming stability of the distribution of misreporting probabilities, the author obtains the confidence sets for the pension plan choice probabilities for individuals in the three survey subsamples for three different periods of time.

6 Conclusion

In this survey we have focused on the recent advances in identification and estimation of nonlinear EIV models with classical measurement errors and nonlinear EIV models with nonclassical measurement errors, as well as some results on partial identification in nonlinear EIV models. We have briefly discussed the applications of various new methods immediately after the methods are introduced. Additional applications using econometric techniques for solving measurement error problems can be found in Carroll, Ruppert, and Stefanski (1995), Bound, Brown, and Mathiowetz (2001) and Moffit and Ridder (to appear).

Due to the lack of time and space, we have not reviewed many papers on measurement errors in details. We have not mentioned any Bayesian approach to measurement error problems. We have not discussed methods to solve measurement errors problems that take advantages of panel data and time series structures; see e.g., Hsiao (1989), Horowitz and Markatou (1996), Dynan (2000) and Parker and Preston (September, 2005) for such applications. We have also not discussed the literature on small noise approximation to assess the effect of measurement errors; see e.g., Chesher (1991), Chesher and Schluter (2002) and Chesher, Dumangane, and Smith (2002).

Despite numerous articles that have been written on the topic of measurement errors in econometrics and statistics over the years, there are still many unsolved important questions. For example, the implications of measurement errors and data contaminations on complex (nonlinear) structural models in labor economics, industrial organization and asset pricing are yet to be understood and studied. Also, it is often the case that not all mismeasured variables are validated in auxiliary data sets; hence how to make use of partial information in validation studies is an important question. Finally, there is relatively little work on the problem of misspecification of various crucial identifying assumptions for nonlinear EIV models.

References

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Amemiya, Y. (1985): "Instrumental variable estimator for the nonlinear errors-in-variables model," *Journal of Econometrics*, 28, 273–290.

Bollinger, C. (1998): "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics*, 16(3), 576–594.

Bonhomme, S., and J. Robin (2006): "Generalized nonparametric deconvolution with an application to earnings dynamics," working paper, University College London.

Bound, J., C. Brown, G. Duncan, and W. Rodgers (1994): "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics*, 12, 345–368.

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement Error in Survey Data," in *Handbook of Econometrics, Vol. 5*, ed. by J. J. Heckman, and E. E. Leamer. North Holland.

Bound, J., and A. Krueger (1991): "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right," *Journal of Labor Economics*, 12, 1–24.

Butucea, C., and M. Taupin (2005): "New M-Estimators in Semiparametric Regression With Errors in Variables," arXiv:math.ST/0511105 v1.

Carroll, R., and P. Hall (1988): "Optimal rates of convergence for deconvolving a density," *Journal of American Statistical Association*, 83, 1184–1186.

Carroll, R., D. Ruppert, C. Crainiceanu, T. Tosteson, and M. Karagas (2004): "Nonlinear and Nonparametric Regression and Instrumental Variables," *Journal of the American Statistical Association*, 99(467), 736–750.

Carroll, R., and M. Wand (1991): "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society*, 53, 573–585.

Carroll, R. J., D. Ruppert, and L. A. Stefanski (1995): *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall.

Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72(2), 343–366.

Chen, X., H. Hong, and A. Tarozzi (2004): "Semiparametric Efficiency in GMM Models Nonclassical Measurement Errors," working paper, Duke University and New York University.

Chen, X., and Y. Hu (2006): "Identification and inference of nonlinear models using two samples with arbitrary measurement errors," Cowles Foundation Discussion Paper No. 1590.

Chesher, A. (1991): "The effect of measurement error," *Biometrika*, 78, 451–462.

Chesher, A., M. Dumangane, and R. Smith (2002): "Duration response measurement error," *Journal of Econometrics*, 111, 169–194.

Chesher, A., and C. Schluter (2002): "Welfare measurement and measurement error," *Review of Economic Studies*, 69, 357–378.

Devereux, P., and G. Tripathi (2005): "Combining datasets to overcome selection caused by censoring and truncation in moment bases models," Working Paper.

Dynan, K. (2000): "Habit Formation in Consumer Preferences: Evidence from Panel Data," *Review of Economic Studies*, 90, 391–406.

Fan, J. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *The Annals of Statistics*, 19(3), 1257–1272.

Fan, J., and Y. Truong (1993): "Nonparametric regression with errors in variables," *Annals of Statistics*, 21, 1900–1925.

Fricsh, R. (1934): *Statistical Confluence Study*. Oslo: University Institute of Economics.

Friedman, M. (1957): *A Theory of the Consumption Function*. Princeton University Press.

Fuller, W. (1987): *Measurement Error Models*. New York: John Wiley & Sons.

Hahn, J. (1998): "On the Role of Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–332.

Hausman, J. (Autumn, 2001): "Mismeasured variables in econometric analysis: problems from the right and problems from the left," *The Journal of Economic Perspectives*, 15, 57–67.

Hausman, J., J. Abrevaya, and F. Scott-Morton (1998): "Misclassification of the Dependent Variable in a Disrete-response Setting," *Journal of Econometrics*, 87, 239–269.

Hausman, J., W. Newey, H. Ichimura, and J. Powell (1991): "Measurement Errors in Polynomial Regression Models," *Journal of Econometrics*, 50, 273–295.

Hausman, J., W. Newey, and J. Powell (1995): "Nonlinear errors in variables estimation of some Engel curves," *Journal of Econometrics*, 65, 205–233.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Hong, H., and E. Tamer (2003): "A Simple Estimator for Nonlinear Error in Variable Models," *Journal of Econometrics*, 117(1), 1–19.

Horowitz, J., and W. Hardle (1994): "Testing a Parametric Model against a Semiparametric Alternative," *Econometric Theory*, 10, 821–848.

Horowitz, J., and C. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.

Horowitz, J., and M. Markatou (1996): "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63, 145–168.

Hsiao, C. (1989): "Identification and estimation of dichotomous latent variables models using panel data," *Review of Economic Studies*, 58, 717–731.

Hu, Y. (2006): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables," Department of Economics, The University of Texas at Austin.

Hu, Y., and G. Ridder (2006): "Estimation of Nonlinear Models with Mismeasured Regressors Using Marginal Informtion," Department of Economics, The University of Texas at Austin and University of Southern California.

Hu, Y., and S. Schennach (2006): "Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions using instruments," Department of Economics, The University of Texas at Austin and University of Chicago.

Ichimura, H., and E. Martinez-Sanchis (2005): "Identification and Estimation of GMM Models by Combining Two Data Sets," Working paper, UCL and CEMMAP.

Lee, L., and J. Sepanski (1995): "Estimation of Linear and Nonlinear Errors-inVariables Models Using Validation Data," *Journal of the American Statistical Association*, 90(429), 130–140.

Lewbel, A. (2006): "Estimation of Average Treatment Effects With Misclassification," forthcoming, Econometrica.

Li, T. (2002): "Robust and consistent estimation of nonlinear errors-in-variables models," *Journal of Econometrics*, 110, 1–26.

Li, T., and C. Hsiao (2004): "Robust estimation of generalized linear models with measurement errors," *Journal of Econometrics*, 118, 51–652.

Li, T., I. Perrigne, and Q. Vuong (2000): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Econometrics*, 98, 129–161.

Li, T., and Q. Vuong (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165.

Linton, O., and Y.-J. Whang (2002): "Nonparametric Estimation with Aggregated Data," *Econometric Theory*, 18, 420–468.

Mahajan, A. (2005): "Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 74(3), 631–665.

Moffit, R., and G. Ridder (to appear): "The econometrics of data combination," in *Handbook of Econometrics, Vol. 6*, ed. by J. J. Heckman, and E. E. Leamer. North Holland.

Molinari, F. (2005): "Partial Identification of Probability Distributions with Misclassified Data," Cornell University, Working Paper.

Newey, W. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

Newey, W. (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors in Variables Models," *Review of Economics and Statistics*, 83(4), 616–627.

Parker, J., and B. Preston (September, 2005): "Precautionary Savings and Consumption Fluctuations," *American Economic Review*, 95(4), 1119–1144.

Powell, J. (1984): "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, pp. 303–325.

Robins, J., S. Mark, and W. Newey (1992): "Estimating exposure effects by modelling the expectation of exposure conditional on confounders," *Biometrics*, 48, 479–95.

Schennach, S. (2004a): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72(1), 33–75.

——— (2004b): "Nonparametric Estimation in the Presence of Measurement Error," *Econometric Theory*, 20, 1046–1093.

Schennach, S. (2006): "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," forthcoming, Econometrica.

Sepanski, J., and R. Carroll (1993): "Semiparametric Quasi-likelihood and Variance Estimation in Measurement Error Models," *Journal of Econometrics*, 58, 223–256.

Taupin, M. L. (2001): "Semiparametric Estimation in the Nonlinear Structural Errors-in-Variables Model," *Annals of Statistics*, 29.

Wang, L. (2004): "Estimation of nonlinear models with Berkson measurement errors," *Annals of Statistics*, 32, 2559–2579.

Wang, L., and C. Hsiao (1995): "Simulation-based semiparametric estimation of nonlinear errors-in-variables models," working paper, University of Southern California.

Wansbeek, T., and E. Meijer (2000): *Measurement Error and Latent Variables in Econometrics.* New York: North Holland.