

Regressions with Berkson errors in covariates- a nonparametric approach

Susanne Schennach

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP22/13

Regressions with Berkson errors in covariates — A nonparametric approach

Susanne M. Schennach*
Brown University

April 13, 2013

Abstract

This paper establishes that so-called instrumental variables enable the identification and the estimation of a fully nonparametric regression model with Berkson-type measurement error in the regressors. An estimator is proposed and proven to be consistent. Its practical performance and feasibility are investigated via Monte Carlo simulations as well as through an epidemiological application investigating the effect of particulate air pollution on respiratory health. These examples illustrate that Berkson errors can clearly not be neglected in nonlinear regression models and that the proposed method represents an effective remedy.

1 Introduction

Many statistical data sets involve covariates X that are error-contaminated versions of their true unobserved counterpart X^* . However, the measurement error often does not fit the classical error structure $X = X^* + \Delta X$ with ΔX independent from X^* . A common occurrence is, in fact, the opposite situation, in which $X^* = X + \Delta X^*$ with ΔX^* independent from X , a situation often referred to as Berkson measurement error (Berkson (1950), Wang (2004), Carroll, Ruppert, Stefanski, and Crainiceanu (2006)). A typical example is an epidemiological study in which an individual's true exposure X^* to some contaminant is not observed, but instead, what is available is the average concentration X of this contaminant in the region where the individual lives. The individual-specific X^* randomly fluctuate around the region average X , resulting in Berkson errors.

Existing approaches to handle data with Berkson measurement error (e.g. Delaigle, Hall, and Qiu (2006), Carroll, Delaigle, and Hall (2007)) unfortunately require the distribution of the measurement error to be known, or to be estimated via validation data, which can be costly, difficult or impossible to collect. (In classical measurement error problems, the distribution of the error can be identified from repeated measurements via a Kotlarski-type

*This work was made possible in part through financial support from the National Science Foundation via grants SES-0752699 and SES-1156347, and through TeraGrid computer resources provided by the University of Texas under grant SES-070003.

equality (Schennach (2004), Li and Vuong (1998)). However, such results do not yet exist for Berkson-type measurement error.) A popular approach to relax the assumption of a fully known distribution of the measurement error is to allow for some adjustable parameters in the distributions of the variables and their relationships, and solve for the parameter values that best reproduce various conditional moments of the observed variables, under the assumption that this solution is unique. This approach has been used, in particular, for polynomial specifications (Huwang and Huang (2000)) and, more recently, for a very wide range of parametric models (Wang (2004), Wang (2007)).

The present paper goes beyond this and provides a formal identification result and a general nonparametric regression method that is consistent in the presence of Berkson errors, without requiring the distribution of the measurement error to be known a priori. Instead, the method relies on the availability of a so-called instrumental variable (e.g., see Chapter 6 in Carroll, Ruppert, Stefanski, and Crainiceanu (2006)) to recover the relationship of interest. For instance, in the epidemiological study of the effect of particulate matter pollution on respiratory health we consider in this paper, suitable instruments could include (i) individual-level measurement of contaminant levels that can even be biased and error-contaminated or (ii) incidence rates of diseases other than the one of interest that are known to be affected by the contaminant in question.

Our estimation method essentially proceeds by representing each of the unknown functions in the model by a truncated series (or a flexible functional form) and by numerically solving for the parameter values that best fits the observable data. Although such an approach is easy to suggest and implement, it is a challenging task to formally establish that such a method is guaranteed to work in general. First, there is no guarantee that the solution (i.e. parameter values that best match the distribution of the observable data) is unique. Second, estimation in the presence of a number of unknown parameters going to infinity with sample size is fraught with convergence questions. Can the postulated series represent the solution asymptotically? Is the parameter space too large to obtain consistency? Is the noise associated with estimating an increasing number of parameters kept under control?

Our solution to these problems is two-fold. First, we target the most difficult obstacle by formally establishing identification conditions under which the regression function and the distribution of all the unobserved variables of the model are uniquely determined by the distribution of the observable variables. A second important aspect of our solution to the Berkson measurement error problem is to exploit the extensive and well-developed literature on nonparametric sieve estimation (e.g., Grenander (1981), Gallant and Nychka (1987), Shen (1997)) to formally address the potential convergence issues that arise when nonparametric unknowns are represented via truncated series with a number of terms that increases with sample size. These theoretical findings are supported by a simulation study and the usefulness of the method is illustrated with an epidemiological application to the effect of particulate matter pollution on respiratory health.

2 Model and Framework

We consider a regression model of the general form

$$Y = g(X^*) + \Delta Y \tag{2.1}$$

$$X^* = X + \Delta X^* \tag{2.2}$$

$$Z = h(X^*) + \Delta Z \tag{2.3}$$

where the function $g(\cdot)$ is the (unknown) relationship of interest between Y , the observed outcome variable and X^* , the *unobserved* true regressor, while ΔY is a disturbance. Information regarding X^* is only available in the form of an observable proxy X contaminated by an error ΔX^* . Equation (2.3) assumes the availability of an instrument Z , related to X^* via an unknown function $h(\cdot)$ and a disturbance ΔZ . Our goal is to estimate the function $g(\cdot)$ in (2.1) nonparametrically and without assuming that the distribution of the measurement error ΔX^* is known. (As by-products, we will also obtain $h(\cdot)$ and the joint distribution of all the unobserved variables.) To this effect, we require the following assumptions, which are very common in the literature focusing on nonlinear models with measurement error (e.g. Carroll, Ruppert, Stefanski, and Crainiceanu (2006), Wang (2004), Hausman, Newey, Ichimura, and Powell (1991), Fan and Truong (1993), Li (2002), Lewbel (1996)).

Assumption 2.1. *The random variables X , ΔX^* , ΔY , ΔZ are mutually independent.*

Note that Assumption 2.1 implies the commonly-made “surrogate assumption” $f_{Y|X,X^*}(y|x,x^*) = f_{Y|X^*}(y|x^*)$, as can be seen by the following sequence of equalities: $f_{Y|X,X^*}(y|x,x^*) = f_{\Delta Y|X,X^*}(y-g(x^*)|x,x^*) = f_{\Delta Y|\Delta X^*,X}(y-g(x^*)|x^*-x,x) = f_{\Delta Y}(y-g(x^*)) = f_{\Delta Y|X^*}(y-g(x^*)|x^*) = f_{Y|X^*}(y|x^*)$.

Assumption 2.2. *The random variables ΔX^* , ΔY , ΔZ are centered (i.e. the model’s restrictions preclude replacing ΔX^* by $\Delta X^* + c$ for some nonzero constant c , and similarly for ΔY and ΔZ ; this includes either zero mean, zero mode or zero median, for instance).*

As our approach relies on the availability of an instrument Z to achieve identification, it is instructive to provide practical examples of suitable instruments in common settings. Although the use of instrumental variables has historically been more prevalent in the econometrics measurement error literature (Hausman, Newey, Ichimura, and Powell (1991); Hausman, Newey, and Powell (1995); Newey (2001); Schennach (2007)), instruments are gathering increasing interest in the statistics literature, especially in the context of measurement error problems (see Chapter 6 entitled “Instrumental Variables” in Carroll, Ruppert, Stefanski, and Crainiceanu (2006) and the numerous references therein).

Note that the instrument Equation (2.3) is entirely analogous to (2.1), the equation generating the main dependent variable. Hence, the instrument is nothing but another observable “effect” caused by X^* via a general nonlinear relationship $h(\cdot)$. Let us consider a few examples, which were inspired by some of the case studies found in Carroll, Ruppert, Stefanski, and Crainiceanu (2006), Wang (2004) and Hyslop and Imbens (2001).

Example 2.1. *Epidemiological studies.*

In these studies, the dependent variable Y is typically a measure of the severity of a disease or condition, while the true regressor X^* is someone’s true but unobserved exposure to some contaminant. The average concentration X of this contaminant in the region where the individual lives is, however, observed. The error on X is Berkson-type because individual-specific X^* typically randomly fluctuate around the region average X . In this setup, multiple plausible instruments are available:

1. A measurement of contaminant concentration in the individual’s house (these would be error-contaminated by classical errors, since the concentration at a given time randomly fluctuates around the time-averaged concentration which would be relevant for the impact on health). Thanks to the flexibility introduced by the function $h(\cdot)$ in (2.3), these measurements can even be biased. They can therefore be made with an inexpensive method (that can be noisy and not even well-calibrated), making it practical to use at the individual level. Hence, it is possible to combine (i) accurate, but expensive, region averages that are not individual-specific (X) and (ii) inexpensive, inaccurate individual-specific measurements (Z) to obtain consistent estimates.
2. Another plausible instrument could be a measure of the severity of another disease or condition that is *known* to be caused by the contaminant. The fact that it is *caused by* the contaminant, introduces an error structure which is consistent with Equation (2.3). Other measurable effects due to the contaminant (e.g., the results of saliva or urine tests for the presence of contaminants, could also serve as instruments. Clearly these measurements are not units of exposure, but the function $h(\cdot)$ can account for this.

Example 2.2. *Experimental studies*

Researchers may wish to study how an effect Y (e.g. the production of some chemical) is related to some imposed external conditions X (e.g., oven or reactor temperature), but the true conditions X^* experienced by the sample of interest may deviate randomly from the imposed conditions (e.g., temperature may not be completely uniform). In this case, an instrument Z could be (i) another “effect” (e.g., the amount of another chemical) that is known to be caused by X^* or (ii) a measurement of X^* that is specific to the sample of interest but that may be very noisy or even biased (e.g. it could be an easier-to-take temperature measurement after the experiment is completed and the sample has partly cooled down.)

Example 2.3. *Self-reported data*

Hyslop and Imbens (2001) have argued that individuals reporting data (e.g. their food intake, or exercise habits) are sometimes aware of the uncertainty in their estimates of X^* and, as a result, try to report an average X over all plausible estimates consistent with the information available to them, thus leading to Berkson-type errors, because the individuals try to make their prediction error independent from their report. In this setting, an instrument Z could be another observable outcome variable Z that is also related to X^* .

3 Identification

We now formally state conditions under which the Berkson measurement error model can be identified with the help of an instrument. Let \mathcal{Y} , \mathcal{X} , \mathcal{X}^* and \mathcal{Z} denote the supports of the distributions of the random variables Y , X , X^* and Z , respectively. We consider Y, X, X^* and Z to be jointly continuously distributed (with $\mathcal{Y} \subset \mathbb{R}^{n_y}$, $\mathcal{X} \subset \mathbb{R}^{n_x}$, $\mathcal{X}^* \subset \mathbb{R}^{n_x}$ and $\mathcal{Z} \subset \mathbb{R}^{n_z}$ with $n_z \geq n_x$). Accordingly, we assume the following.

Assumption 3.1. *The random variables Y, X, X^*, Z admit a bounded joint density with respect to the Lebesgue measure on $\mathcal{Y} \times \mathcal{X} \times \mathcal{X}^* \times \mathcal{Z}$. All marginal and conditional densities are also defined and bounded.*

We use the notation $f_A(a)$ and $f_{A|B}(a|b)$ to denote the density of the random variable A and the density of A conditional on B , respectively. Lower case letters denote specific values of the corresponding upper case random variables. Next, as in many treatments of errors-in-variables models (Carroll, Ruppert, Stefanski, and Crainiceanu (2006), Fan and Truong (1993), Li and Vuong (1998), Li (2002), Schennach (2004), Schennach (2007)), we require various characteristic functions to be nonvanishing. We also place regularity constraints on the two regression functions of the model.

Assumption 3.2. *For all $\zeta \in \mathbb{R}^{n_z}$, $E[\exp(\mathbf{i}\zeta \cdot \Delta Z)] \neq 0$ and for all $\xi \in \mathbb{R}^{n_x}$, $E[\exp(\mathbf{i}\xi \cdot \Delta X^*)] \neq 0$ (where $\mathbf{i} = \sqrt{-1}$).*

Assumption 3.3. *$g : \mathcal{X}^* \mapsto \mathcal{Y}$ and $h : \mathcal{X}^* \mapsto \mathcal{Z}$ are one-to-one (but not necessarily onto).*

Assumption 3.4. *h is continuous.*

Assumption 3.3 is somewhat restrictive when X^* has a dimension larger or equal to the ones of Y (or Z). Fortunately, it is often possible to eliminate this problem by re-defining Y (and Z) to be a vector containing auxiliary variables in addition to the outcome of interest, in order to allow for enough variation in Y (and Z) to satisfy Assumption 3.3. Each of these additional variables need not be part of the relationship of interest per se, but does need to be affected by X^* in some way. In that sense, such auxiliary variables would also be a type of ‘‘instrument’’. Our main identification result can then be stated as follows. (Note that the theorem also holds upon conditioning on an observed variable W , so that additional, correctly measured, regressors can be straightforwardly included.)

Theorem 3.1. *Under Assumptions 2.1-3.4, given the true observed conditional density $f_{Y,Z|X}$, the solution $(g, h, f_{\Delta Z}, f_{\Delta Y}, f_{\Delta X^*})$ to the functional equation*

$$f_{Y,Z|X}(y, z|x) = \int f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - g(x^*)) f_{\Delta X^*}(x^* - x) dx^* \quad (3.1)$$

for all $y \in \mathcal{Y}$, $x \in \mathcal{X}$, $z \in \mathcal{Z}$ is unique (up to differences on sets of null probability measure). A similar uniqueness result holds for the solution $(g, h, f_{\Delta Z}, f_{\Delta Y}, f_{\Delta X^}, f_X)$ to*

$$f_{Y,Z,X}(y, z, x) = f_X(x) \int f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - g(x^*)) f_{\Delta X^*}(x^* - x) dx^*. \quad (3.2)$$

Establishing this result demands techniques radically different from existing treatment of Berkson error models, such as the spectral decomposition of linear operators (see Carrasco, Florens, and Renault (2005) for a review), which are emerging as powerful alternatives to the ubiquitous deconvolution techniques that are typically applied in classical measurement error problems. The proof can be found in Appendix A and can be outlined as follows. Assumption 2.1 lets us obtain the following integral equation relating the joint densities of the observable variables to the joint densities of the unobservable variables:

$$f_{Y,Z|X}(y, z|x) = \int f_{Z|X^*}(z|x^*) f_{Y|X^*}(y|x^*) f_{X^*|X}(x^*|x) dx^* \quad (3.3)$$

from which Equation (3.1) follows directly. Uniqueness of the solution is then shown as follows. Equation (3.3) defines the following operator equivalence relationship:

$$F_{y;Z|X} = F_{Z|X^*} D_{y;X^*} F_{X^*|X}, \quad (3.4)$$

where we have introduced the following operators:

$$\begin{aligned} [F_{y;Z|X}r](z) &= \int f_{Y,Z|X}(y, z|x) r(x) dx & [F_{Z|X^*}r](z) &= \int f_{Z|X^*}(z|x^*) r(x^*) dx^* \\ [F_{Z|X}r](z) &= \int f_{Z|X}(z|x) r(x) dx & [D_{y;X^*}r](x^*) &= f_{Y|X^*}(y|x^*) r(x^*) \\ [F_{X^*|X}r](x^*) &= \int f_{X^*|X}(x^*|x) r(x) dx. \end{aligned} \quad (3.5)$$

for some sufficiently regular but otherwise arbitrary function r . Note that, in the above definitions, y is viewed as a parameter (the operators do not act on it) and that $D_{y;X^*}$ is the operator equivalent of a diagonal matrix. Next, we note that the equivalence $F_{Z|X} = F_{Z|X^*} F_{X^*|X}$ also holds (e.g., by integration of (3.4) over all $y \in \mathcal{Y}$). We can then isolate $F_{X^*|X}$

$$F_{X^*|X} = F_{Z|X^*}^{-1} F_{Z|X} \quad (3.6)$$

and substitute the result into (3.4) to yield, after rearrangements:

$$F_{y;Z|X} F_{Z|X}^{-1} = F_{Z|X^*} D_{y;X^*} F_{Z|X^*}^{-1}, \quad (3.7)$$

where all inverses can be shown to exist over suitable domains under our assumptions. Equation (3.7) states that the operator $F_{y;Z|X} F_{Z|X}^{-1}$ admits a spectral decomposition. The operator to be “diagonalized” is defined in terms of observable densities, while the resulting eigenvalues $f_{Y|X^*}(y|x^*)$ (contained in $D_{y;X^*}$) and eigenfunctions $f_{Z|X^*}(\cdot|x^*)$ (contained in $F_{Z|X^*}$) provide the unobserved densities of interest.

A few more steps are required to ensure uniqueness of this decomposition, which we now briefly outline. One needs to (i) invoke a powerful uniqueness result regarding spectral decompositions (Theorem XV 4.5 in Dunford and Schwartz (1971)), (ii) exploit the fact that densities integrate to one to fix the scale of the eigenfunctions, (iii) handle degenerate eigenvalues and (iv) uniquely determine the ordering and indexing of the eigenvalues and eigenfunctions. This last, and perhaps most difficult, step, addresses the issue that both $f_{Z|X^*}(\cdot|x^*)$ and $f_{Z|X^*}(\cdot|S(x^*))$, for some one-to-one function S , are equally valid ways to state the eigenfunctions that nevertheless result in different operators $F_{Z|X^*}$. To resolve this ambiguity, we note that for any possible operator $F_{Z|X^*}$ satisfying (3.7), there exist a unique

corresponding operator $F_{X^*|X}$, via Equation (3.6). However, only one choice of $F_{Z|X^*}$ leads to an operator $F_{X^*|X}$ whose kernel $f_{X^*|X}(x^*|x)$ satisfies Assumption 2.2. Hence, $f_{X^*|X}(x^*|x)$, $f_{Y|X^*}(y|x^*)$ and $f_{Z|X^*}(z|x^*)$ are identified, from which the functions $f_{\Delta Z}$, $f_{\Delta Y}$, $f_{\Delta X^*}$, h and g can be recovered by exploiting the centering restrictions on ΔX^* , ΔY and ΔZ .

An operator approach has recently been proposed to address certain types of nonclassical measurement error problems (Hu and Schennach (2008)), but under assumptions that rule out Berkson-type measurement errors: It should be emphasized that, despite the use of operator decomposition techniques similar to the ones found in Hu and Schennach (2008) (hereafter HS), it is impossible to simply use their results to identify the Berkson measurement error model considered here, for a number of reasons. First, the key condition (Assumption 5 in HS) that the distribution of the mismeasured regressor X given the true regressor X^* is “centered” around X^* does not hold for Berkson errors. Consider the simple case where the Berkson measurement error is normally distributed and so are the true and mismeasured regressors. The distribution of X given $X^* = x^*$ is a normal centered at $x^* \sigma_x^2 / (\sigma_x^2 + \sigma_{\Delta x^*}^2)$. Hence, there is absolutely no reasonable measure of location (mean, mode, median, etc.) that would yield the appropriate centering at x^* that is needed in Assumption 5 of HS. In addition, one cannot simply replace the assumption of centering of X given X^* (as in HS) by a centering of X^* given X (as would be required for Berkson errors) and hope that Theorem 1 in HS remains valid. HS exploit the fact that, in a conditional density, there is no Jacobian term associated with a change of variable in a conditioning variable (here X^*). However, with Berkson errors, the corresponding change of variable would not take place in the conditional variables, and a Jacobian term would necessarily appear, which makes the approach used in HS fundamentally inapplicable to the Berkson case. Solving this problem involves (i) using a different operator decomposition than in HS and (ii) using a completely different approach for “centering” the mismeasured variable.

A referee suggested an alternative argument (formalized in the Appendix) that makes a more direct connection with Theorem 1 in HS but under the additional assumption that Z and X^* have the same dimension. Such an assumption is rather restrictive because it will often result in the assumption that $h(\cdot)$ is one-to-one (Assumption 3.3) being violated. For instance, if X^* is scalar and we have access to two instruments Z_1 and Z_2 such that neither $E[Z_1|X^*]$ nor $E[Z_2|X^*]$ are strictly monotone, then $h(\cdot)$ is not one-to-one for either instrument used in isolation. However, the mapping $X^* \mapsto (E[Z_1|X^*], E[Z_2|X^*])$ will typically be one-to-one, except for really exceptional cases. Hence, allowing for the dimensions of X^* and Z to differ is important. Nevertheless, even assuming away this problem, such an approach still requires a different technique for centering X^* than the one used in HS. That said, both HS and the current paper rely on operator spectral decomposition as an alternative to conventional convolution/deconvolution techniques, and it appears likely that these new techniques will find applications in a number of other measurement error models.

Observe that our identification result is also useful in a parametric and semi-parametric context, as it provides the confidence that, under simple conditions, the model is identified. Rank conditions that would need to be verified on a case-by-case basis in any given parametric model are automatically implied by our identification results in a wide class of models. Also, although X is allowed to be random throughout, considering X to be fixed poses no particular difficulty, since Equation (3.1) provides a valid conditional likelihood function in that case.

As discussed in Appendices D and E, a number of extensions of the method are possible:

(i) Relaxing the independence between X and ΔX^* to allow for some heteroskedasticity in the measurement error and (ii) combining classical and Berkson errors, a possibility considered in, e.g., Mallick, Hoffman, and Carroll (2002), Carroll, Delaigle, and Hall (2007), Stram, Huberman, and Wu (2002) and Hyslop and Imbens (2001). It can also be shown that some extensions are not plausible, such as assuming that both the measurement equation (2.2) and the instrument equation (2.3) have a Berkson error structure (see Appendix D).

4 Estimation

A natural way to obtain a nonparametric estimator of the model is to substitute truncated series approximations into (3.1) or (3.2) for each of the unknown functions and construct a log likelihood function to be maximized numerically with respect to all coefficients of the series (e.g. Shen (1997)). Such sieve-based estimators have recently found applications in a variety of measurement error problems (e.g., Newey (2001), Mahajan (2006), Hu and Schennach (2008), Carroll, Chen, and Hu (2010), among others). Below we first define our estimator before establishing its consistency.

We represent the regression functions $g(\cdot)$ and $h(\cdot)$ as

$$\hat{m}^{(K_m)}(x^*, \beta_m^{(K_m)}) = \sum_{k=1}^{K_m} \beta_{m,k}^{(K_m)} q_k^{(K_m)}(x^*) \text{ for } m = g, h \quad (4.1)$$

where $q_k^{(K_m)}(x^*)$ is some sequence (indexed by the truncation parameters K_m) of progressively larger sets of basis functions indexed by $k = 1, \dots, K_m$ while $\beta_m^{(K_m)} = (\beta_{m,1}^{(K_m)}, \dots, \beta_{m,K}^{(K_m)})$ is a vector of coefficients to be determined. The $q_k^{(K_m)}(x^*)$ could be some power series, trigonometric series, orthogonal polynomials, wavelets or splines, for instance. The double indexing by k and K_m is useful to allow for splines, where changing the number of knots modifies all the basis functions.

A similar expansion in terms of basis functions $p_k^{(K_V)}(v)$ (with truncation parameter K_V) is used for the density of each disturbance $V = \Delta Z, \Delta Y, \Delta X^*$:

$$\hat{f}_V^{(K_V)}(v, \theta_V^{(K_V)}) = \frac{1}{\theta_{V,0}^{(K_V)}} \phi_0(v/\theta_{V,0}^{(K_V)}) \sum_{k=1}^{K_V} \theta_{V,k}^{(K_V)} p_k^{(K_V)}(v), \quad (4.2)$$

where $\theta_V^{(K_V)} = (\theta_{V,0}^{(K_V)}, \dots, \theta_{V,K}^{(K_V)})$ is a vector of coefficients to be determined and $\phi_0(\cdot)$ is a user-specified “baseline” function. The “baseline” function is convenient to reduce the number of terms needed in the expansion, when the approximate general shape of the density is known. It is not strictly needed, however, and can be set to 1. Either way, the method is fully nonparametric. A convenient choice of basis (see Gallant and Nychka (1987)) is to take $\phi_0(\cdot)$ to be a Gaussian and $p_k^{(K_V)}(v) = v^{k-1}$ for any K_V .

An important distinction with the functions $g(\cdot)$ and $h(\cdot)$ is that some constraints have to be imposed on the densities. One constraint is needed to ensure centering (Assumption 2.2):

$$\sum_{k=1}^{K_V} \theta_{V,k}^{(K_V)} C_{V,c,k}^{(K_V)} = 0,$$

where, for some user-specified function $c_V(v)$, we define

$$C_{V,c,k}^{(K_V)} = \int c_V(v) \frac{1}{\theta_{V,0}^{(K_V)}} \phi_0\left(\frac{v}{\theta_{V,0}^{(K_V)}}\right) p_k^{(K_V)}(v) dv.$$

For instance, to impose zero mean on the disturbance V , let $c_V(v) = v$. To impose zero median, let $c_V(v) = \mathbf{1}(v \leq 0) - 1/2$, where $\mathbf{1}(\cdot)$ denotes an indicator function, while to impose zero mode, let $c_V(v) = -\delta^{(1)}(v)$ (a delta function derivative, in a slight abuse of notation). Another constraint is needed to ensure unit total probability: $\sum_{k=1}^{K_V} \theta_{V,k}^{(K_V)} C_{V,1,k}^{(K_V)} = 1$. Note that both types of constraints exhibit the computationally convenient property of being linear in the unknown coefficients.

Given the above definitions, we can define an estimator of all unknown functions based on a sample $(X_i, Y_i, Z_i)_{i=1}^n$ and Equation (3.1) (a corresponding estimator based on Equation (3.2) can be derived analogously). Let $\hat{\beta}_g^{(K_g)}, \hat{\beta}_h^{(K_h)}, \hat{\theta}_{\Delta X^*}^{(K_V)}, \hat{\theta}_{\Delta Y}^{(K_V)}, \hat{\theta}_{\Delta Z}^{(K_V)}$ denote the minimizer of the sample log likelihood

$$\frac{1}{n} \sum_{i=1}^n \ln \hat{f}_{Y,Z|X}(Y_i, Z_i | X_i) \quad (4.3)$$

where

$$\begin{aligned} \hat{f}_{Y,Z|X}(y, z | x) &= \int \hat{f}_{\Delta Z}^{(K_{\Delta Z})}\left(z - \hat{h}^{(K_h)}\left(x^*, \hat{\beta}_h^{(K_h)}\right), \theta_{\Delta Z}^{(K_{\Delta Z})}\right) \times \\ &\quad \hat{f}_{\Delta Y}^{(K_{\Delta Y})}\left(y - \hat{g}^{(K_g)}\left(x^*, \hat{\beta}_g^{(K_g)}\right), \theta_{\Delta Y}^{(K_{\Delta Y})}\right) \hat{f}_{\Delta X^*}^{(K_{\Delta X^*})}\left(x^* - x, \theta_{\Delta X^*}^{(K_{\Delta X^*})}\right) dx^*, \end{aligned}$$

subject to

$$\sum_{k=1}^{K_V} \theta_{V,k}^{(K_V)} C_{V,1,k}^{(K_V)} = 1 \text{ and } \sum_{k=1}^{K_V} \theta_{V,k}^{(K_V)} C_{V,c,k}^{(K_V)} = 0 \quad (4.4)$$

for $V = \Delta Z, \Delta Y, \Delta X^*$ and subject to technical regularity constraints to be defined below. Estimators are then given by

$$\begin{aligned} \hat{g}(x^*) &= \hat{g}^{(K_g)}\left(x^*; \hat{\beta}_g^{(K_g)}\right), \quad \hat{h}(x^*) = \hat{h}^{(K_h)}\left(x^*; \hat{\beta}_h^{(K_h)}\right) \\ \hat{f}_V(v) &= \hat{f}_V^{(K_V)}\left(v, \hat{\theta}_V^{(K_V)}\right) \text{ for } V = \Delta X^*, \Delta Y, \Delta Z. \end{aligned} \quad (4.5)$$

This type of estimator falls within the very general class of sieve nonparametric maximum likelihood estimators (MLE), whose asymptotic theory has received considerable attention over the last few decades (e.g. Grenander (1981), Gallant and Nychka (1987), Shen (1997)). Here, we parallel the treatment of Gallant and Nychka (1987) and Newey and Powell (2003) to establish the consistency of the above procedure. Although the consistency of sieve-type estimator has been previously established in very general settings under some high-level assumptions, our contribution is to provide very primitive sufficient conditions for consistency for the class of models considered here.

We first need to define the set in which the densities of interest reside. The formal proof of consistency of the estimator requires this set to be compact, although this requirement appears to have little impact in practice. In essence, compactness is helpful to rule out very

extreme but rare events associated with very poor estimates. It is a standard regularity condition (see, e.g. Gallant and Nychka (1987), Newey and Powell (2003), Newey (2001)). A well-known type of infinite-dimensional but compact sets are those generated via boundedness and Lipschitz constraints in an \mathcal{L}_∞ space. Here, we use a weighted Lipschitz constraint in order to allow for densities supported on an unbounded set, while still maintaining compactness (our treatment can be straightforwardly adapted to cover the simpler case where the variables are supported on finite intervals). Following Gallant and Nychka (1987), we enforce restrictions that avoid too rapid divergences in the log likelihood.

Definition 4.1. Let $\|f\| = \sup_{v \in \mathbb{R}} |f(v)|$. Let B be finite and strictly positive. Let $f'_+(v)$ be strictly positive and bounded function that is decreasing in $|v|$, symmetric about $v = 0$ and such that $\int_{-\infty}^{\infty} f'_+(v) dv < \infty$. Let $\mathcal{S} = \{f : \mathbb{R} \mapsto [-B, B] \text{ such that } |\partial^\lambda f(v) / \partial v^\lambda| \leq f'_+(v)\}$. Let $f_-(v)$ and $f_+(v)$ be strictly positive and bounded functions with $f_-(v)$ decreasing in $|v|$ and $\int_{-\infty}^{\infty} f_+(v) dv < \infty$. Let $\mathcal{F} = \{f \in \mathcal{S} : f_-(v) \leq f(v) \leq f_+(v)\}$.

We also define suitable norms and sets for the regression functions. Here, we need to allow for functions that diverge to infinity at controlled rates towards infinite values of their argument. In analogy with any existing global measure of expected error, we also use a norm that downweights errors in the tails, which is consistent with the fact that the tails of a nonparametric regression function are always estimated with more noise, since there are fewer datapoints there.

Definition 4.2. Let $\omega : \mathbb{R} \mapsto \mathbb{R}^+$ be some given strictly positive, bounded and differentiable weighting function. For any function $g : \mathbb{R} \mapsto \mathbb{R}$, let $\|g\|_\omega = \|\omega g\|$ where $\omega g(v) \equiv g(v) \omega(v)$. Let $\mathcal{G} = \{g : \omega g \in \mathcal{S} \text{ and } |g(v)| \leq g_+(v)\}$ where $g_+(v)$ is a given positive function that is increasing in $|v|$ and symmetric about $v = 0$.

We can now state the regularity conditions needed.

Assumption 4.1. The observed data (X_i, Y_i, Z_i) are independent and identically distributed across $i = 1, 2, \dots$

Assumption 4.2. We have $f_{\Delta X^*}, f_{\Delta Y}, f_{\Delta Z} \in \mathcal{F}$ and $g, h \in \mathcal{G}$.

Assumption 4.3. The set of functions representable as series (4.2) and (4.1) are, respectively, dense in \mathcal{F} (in the norm $\|\cdot\|$) and \mathcal{G} (in the norm $\|\cdot\|_\omega$).

Denseness results for numerous types of series are readily available in the literature (e.g. Newey (1997), Gallant and Nychka (1987)). Although such results are sometimes phrased in a mean square-type norm rather than the sup norm used here, Lemma 4.1 below (proven in Appendix B) establishes that, within the sets \mathcal{F} and \mathcal{G} , denseness in a mean square norm implies denseness in the norms we use.

Lemma 4.1. Let $\{f_n\}$ be a sequence in \mathcal{F} . Then $\int |f_n(v)|^2 dv \rightarrow 0$ implies $\|f_n\| \rightarrow 0$ (for \mathcal{F} and $\|\cdot\|$ as in Definition 4.1).

We also need standard boundedness and dominance conditions.

Assumption 4.4. For any $x \in \mathbb{R}$, $\int (\omega(x^*))^{-1} f_+(x^* - x) dx^* < \infty$ for ω and f_+ as in Definitions 4.2 and 4.1, respectively.

Assumption 4.5. There exists $b > 0$ such that $E[|\ln(f_-(X, Y, Z))|] < \infty$, where

$$f_-(x, y, z) \equiv 2bf_-(b) f_-(|y| + (g_+(|x| + b))) f_-(|z| + (g_+(|x| + b)))$$

for f_- and g_+ as in Definitions 4.1 and 4.2, respectively.

We can then state our consistency result (proven in Appendix B):

Theorem 4.1. Under Assumptions 3.1-4.5, if $K_V \xrightarrow{p} \infty$, for $V = h, g, \Delta X^*, \Delta Y, \Delta Z$, the estimators given by (4.5) evaluated at the minimizer of (4.3) subject to (4.4), $\hat{f}_{\Delta X^*}, \hat{f}_{\Delta Y}, \hat{f}_{\Delta Z} \in \mathcal{F}$, and $\hat{g}, \hat{h} \in \mathcal{G}$ and satisfying Assumption 4.4 are such that $\|\hat{g} - g^*\|_\omega \xrightarrow{p} 0$, $\|\hat{h} - h^*\|_\omega \xrightarrow{p} 0$, $\|\hat{f}_{\Delta X^*} - f_{\Delta X^*}^*\| \xrightarrow{p} 0$, $\|\hat{f}_{\Delta Y} - f_{\Delta Y}^*\| \xrightarrow{p} 0$, $\|\hat{f}_{\Delta Z} - f_{\Delta Z}^*\| \xrightarrow{p} 0$, where the starred quantities denote the true values (i.e. the unique solution to (3.1)).

The practical implementation of the above approach necessitates the selection of the number of terms K_V in each of the approximating series. Theorem 4.1 allows for a data-driven selection of the K_V , since K_V is allowed to be random. To select the K_V , one can employ the bootstrap cross-validation model selection method based on the Kullback-Leibler (KL) criterion, shown by van der Laan, Dudoit, and Keles (2004) to be consistent even when the number of candidate models grows to infinity with sample size (as it is here). In this method, a fraction p of the sample is excluded at random and the remaining $1 - p$ fraction is used to estimate the model parameters with given numbers ($K_{\Delta X^*}, K_{\Delta Y}, K_{\Delta Z}, K_g, K_h$) of terms in the corresponding series. The likelihood (or KL criterion) is then evaluated using the excluded fraction p at the value of the estimated parameters found in the previous step. The process is repeated many times with different random partitions of the sample into fractions p and $(1 - p)$, to obtain an average KL criterion with a sufficiently small variance (which can be estimated from the KL criterion of each random partitions). This procedure is carried out for various trial choices of ($K_{\Delta X^*}, K_{\Delta Y}, K_{\Delta Z}, K_g, K_h$) and the choice that yields the largest likelihood is selected. This method is consistent asymptotically (as sample size $n \rightarrow \infty$) as $np \rightarrow \infty$ and $p \rightarrow 0$ and under some mild technical regularity conditions stated in van der Laan, Dudoit, and Keles (2004).

Our nonparametric approach nests parametric and semiparametric models. These subcases can be easily implemented by replacing some, or all, of the nonparametric series approximations by suitable parametric models. It is possible to obtain convergence rates and limiting distribution results, along the lines of Shen (1997) or Hu and Schennach (2008), although we do not do so here due to space limitations (stating suitable regularity conditions, even in high-level form, is rather involved, as seen in the Supplementary material of Hu and Schennach (2008), which covers a related but different measurement error model). It is, however, important to point out one important property. Sieve nonparametric MLE is optimal in the following sense: Under suitable regularity conditions, any sufficiently regular semiparametric functional of the nonparametric sieve MLE estimates is asymptotically normal and root n consistent and reaches the semiparametric efficiency bound for that functional (see Theorem 4 in Shen (1997)). This notion of optimality is a natural nonparametric generalization of the well-known efficiency of parametric maximum likelihood.

5 Simulations Study

We now investigate the practical performance and feasibility of the proposed estimator via a simulation example purposely chosen to be a difficult case. The data is generated as follows. The distribution of X is a uniform distribution over $[-1, 1]$ (implying a standard deviation of 0.58). We consider a thick-tailed t distribution with 6 degrees of freedom scaled by 0.5 as the distribution of ΔX^* . The standard deviation of the error ΔX^* is almost identical to the one of the “signal” X , thus making this estimation problem exceedingly difficult. The distribution of ΔY is a logistic scaled by 0.125 while the distribution of ΔZ is a t distribution with 6 degrees of freedom scaled by 0.25. The regression function has the form

$$g(x^*) = |x^*| x^*, \quad (5.1)$$

which is only finitely many times differentiable, thus limiting the convergence rate of its series estimator in the measurement-error-robust estimator (the naive estimator would be less affected since it would “see” a smoothed version of this function). The instrument equation has a specification that is strictly convex and therefore tends to exacerbate the bias in many nonparametric estimators:

$$h(x^*) = \ln(1 + \exp(2x^*)).$$

A total of 100 independent samples, each containing 500 observations, were generated as above and fed into our estimator. For estimation purposes, the functions $g(\cdot)$ and $h(\cdot)$ are both represented by polynomials while the densities of ΔX^* , ΔY and ΔZ are represented by a Gaussian multiplied by a polynomial (following Gallant and Nychka (1987), who establish that these choices satisfy a suitable denseness condition). The Gaussian is centered at the origin but its width is left as a parameter to be estimated. Note that the functional forms considered are not trivially nested within the space spanned by the truncated sieve approximation. This was an intentional choice aimed at properly accounting for the nonparametric nature of the problem (in which the researcher never has the fortune of selecting a truncated sieve fitting the true model exactly).

The integral in Equation (3.1) is evaluated numerically by discretizing the integral as a sum over the range $[-3, 3]$ in intervals of 0.05. Naive least-squares estimators ignoring measurement error (i.e. least-squares regressions of Y on X and of Z on X) were used as a starting point for the numerical sieve optimization of the g and h functions, while the variances of the corresponding residuals were used to construct an initial Gaussian guess for the optimization of all the error distributions. The simplex method due to Nelder and Mead (1965) (also known as “amoeba”) was used to carry out the numerical optimization of the log likelihood (4.3) with respect to all the parameters $\theta_V^{(K_V)}$ for $V = \Delta X^*, \Delta Y, \Delta Z$ and $\beta_m^{(K_m)}$ for $m = g, h$ simultaneously. The constraints that the estimated densities and regression functions lie, respectively, in the sets \mathcal{F} and \mathcal{G} of the form given in Definitions 4.1 and 4.2 are implied by bounds on the magnitude of the sieve coefficients $\theta_{V,k}^{(K_V)}$ and $\beta_{m,k}^{(K_m)}$ in (4.2) and (4.1). Such constraints are easy to impose within the simplex optimization method: parameter changes that would yield violations of the bounds are simply rejected (effectively assigned an “infinite” value) — the simplex optimization method easily accommodates such extreme behavior in the objective function, since it does not rely on derivatives. However,

we found that these constraints are rarely binding in practice, unless the number of terms K_V in the expansions is large (Gallant and Nychka (1987) reports a similar observation). Such large values of K_V tend to be naturally ruled out via our data-driven selection method of the number of terms.

To select the number of terms in the approximating series for a given sample, we use the “bootstrap cross-validation” method described in Section 4 with a fraction $p = 1/8$ and 100 bootstrap replications. Trial values of the number of free parameters (not counting parameters uniquely determined by zero mean and unit area constraints) in the series representing $f_{\Delta X^*}, f_{\Delta Y}, f_{\Delta Z}$ each span the set $\{1, 2, 3, 4\}$ while for g, h each span the set $\{4, 5, 6, 7\}$. The optimal numbers of parameters (kept constant during the replications) were found to be $f_{\Delta X^*} : 3; f_{\Delta Y} : 3; f_{\Delta Z} : 3; g : 6; h : 6$.

Figure 1 summarizes the result of these simulations, where a naive nonparametric series least-squares estimator ignoring measurement error (i.e. least-square regressions of Y on X and of Z on X) with the same number of sieve terms is also shown for comparison. The reliability of the method can be appreciated by noting how closely the median of the replicated measurement-error-robust estimates matches the true model, while the naive estimator ignoring the presence of measurement error is considerably more biased, even missing the fact that the true regression function is nearly flat in the middle section and instead producing a very misleading linear shape despite the strong nonlinearity of the true model. In fact, unlike the proposed estimator, the naive estimator is so significantly biased that any type of hypothesis test based on it would exhibit completely misleading confidence levels: The true model curves (for g and h) almost always lies beyond the 95% or 5% percentiles of the estimator distribution.

Overall, the proposed measurement-error-robust estimator exhibits low variability and low bias at the reasonable sample size of 500. The bias is not exactly zero in a finite sample because our estimator is a nonlinear functional of sample averages and because the sieve approximation necessarily has a limited accuracy in a finite sample. Nevertheless, the fact that our estimator performs so well in the presence of measurement error of such large magnitude is a strong indication of its practical usefulness. This behavior is not specific to this model — we have tested the method in other simulation settings (see Appendix C).

6 Application

Numerous studies have sought to quantify the effect of air pollution on respiratory health (e.g. Dockery, Pope, Xu, et al. (1993)). Specifically, there is a growing concern regarding the effect of small particulate matter (Pope, Thun, Namboodiri, et al. (1995), Samet, Dominici, Curriero, et al. (2000)). A key difficulty with such studies is that air quality monitors are not necessarily located near the subjects being affected by air pollution, implying that the main regressor of interest is mismeasured.

Our approach to this question relies on very comprehensive country-wide data collected by Environment Protection Agency (EPA) and the Center for Disease Control (CDC) in the United States. Pollution levels are taken from EPA’s Monitor Values Report - Criteria Air Pollutants database for year 2005. EPA’s data provides point measurements of the particulate matter levels (we focus on so-called 95th percentile level of PM2.5 particles, those

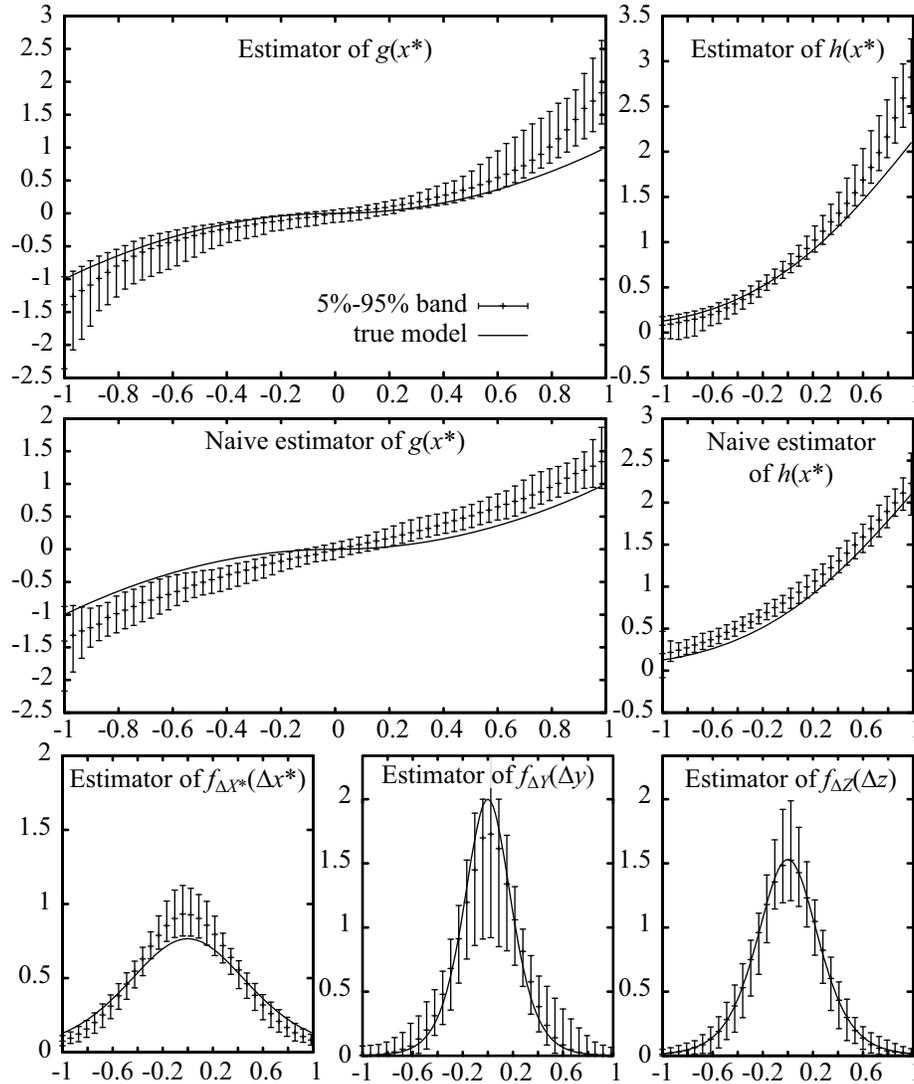


Figure 1: Simulation study of the practical performance of the proposed measurement-error-robust estimator in comparison with a “naive” nonparametric polynomial series least-square estimator that ignores the presence of measurement error. In each plot, the pointwise 90% confidence band of the estimator simulated over 100 replications is shown as error bars.

having less than 2.5 micrometers in diameter) at various monitoring stations throughout the United States, from which we construct state-averaged pollution levels (our X variable, measured in μg of particles per m^3). We do so because pollution data is only available for a small fraction of counties and even where it is available, the nature of its measurement error is complex (it could be a mixture of classical and Berkson errors). By constructing state-level averages, we average out the randomness in monitor measurements while leaving the randomness in the individual exposure untouched, thus obtaining a valid Berkson error-contaminated estimate of the pollution level experienced by individuals from each state, whether they live in a county with a monitoring station or not. Each individual faces an exposure equal to the state average plus an unknown random noise due to his/her precise geographic whereabouts and lifestyle.

Health data is obtained from the publicly available “CDC Wonder” database entitled “Mortality - underlying cause of death ” for year 2005. To measure respiratory health, we use data on causes of death, which offers the advantage that it is very comprehensive and accurate (medical professionals are required to collect it and there is no reliance on voluntary surveys). One limit to the completeness of the data is that, for some counties, the data is “suppressed” (for privacy reasons) or labelled as “unreliable” by the CDC and were therefore omitted from our sample. Our dependent variable of interest (Y) is the rate (per 10,000) of death due to “chronic lower respiratory diseases” (e.g. asthma, bronchitis, emphysema), while our instrument (Z) is the rate (per 10,000) of death resulting from “lung diseases due to external agents” (e.g. pneumoconiosis due to organic or inorganic dust, coalworker’s pneumoconiosis). The rationale is to use, as an instrument, a variable that is clearly expected to be affected by pollution levels. This variable indirectly provides information regarding the true level of pollution, so that the effect of pollution (if any) on the variable of interest can be more accurately assessed. We employ county-level data on causes of death because they are readily available without concerns for patient privacy issues. Moreover, the CDC provides age-corrected death rates, thus correcting for demographic differences between counties. We construct our sample by matching mortality data via counties and matching pollution data via states, resulting in 1305 observations over as many counties and covering all 51 states. A limitation of our approach is that it does not control for other possible confounding effects, e.g., if the proportion of smokers differs between industrial and non-industrial cities. However, such a limitation is common in studies of this kind (as noted in Dockery, Pope, Xu, et al. (1993)).

We use the same types of sieves and computational methods as in the simulation example and select the number of terms using the “bootstrap cross-validation” method described in Section 4 with a fraction $p = 1/8$ and 100 bootstrap replications. Trials values of the number of free parameters in the series representing $f_{\Delta X^*}, f_{\Delta Y}, f_{\Delta Z}$ span the range $\{1, 2, 3\}$ while trial values of the number of terms in the series representing g and h span the range $\{2, 3, 4\}$ (increasing any one of the K_V beyond that range resulted in clearly worse performances). The optimal numbers of free parameters (not counting parameters uniquely determined by zero mean and unit area constraints) were found to be $f_{\Delta X^*} : 2; f_{\Delta Y} : 3; f_{\Delta Z} : 1; g : 4; h : 3$. Pointwise 90% confidence bands around the nonparametric estimates were obtained using the standard bootstrap (see, e.g., Gine and Zinn (1990) for general conditions justifying its use) with 100 replications.

Results are shown in Figure 2. A few observations are in order. First, our measurement

error-robust estimator is perfectly able to detect a clear monotone relationship between Y and X^* and between Z and X^* with useful confidence bands, despite the use of a fully nonparametric approach. Second, although the distribution of the measurement error is difficult to estimate (as reflected by the wide confidence bands), the impact of this uncertainty on the main function of interest ($g(x^*)$) is fortunately very limited. The 90% confidence bands indicate that the presence of substantial measurement error is consistent with the data: The measurement error is of the order of $10 \mu\text{g}/\text{m}^3$, whereas the observed X roughly ranges from 10 to $40 \mu\text{g}/\text{m}^3$. Third, the distribution of ΔY exhibits nonnegligible asymmetry, thus illustrating the drawbacks of methods merely assuming normality of all the error terms. In contrast, the distributions of ΔX^* and ΔZ are apparently very close to symmetric (this is a conclusion of the formal model selection procedure, not an assumption).

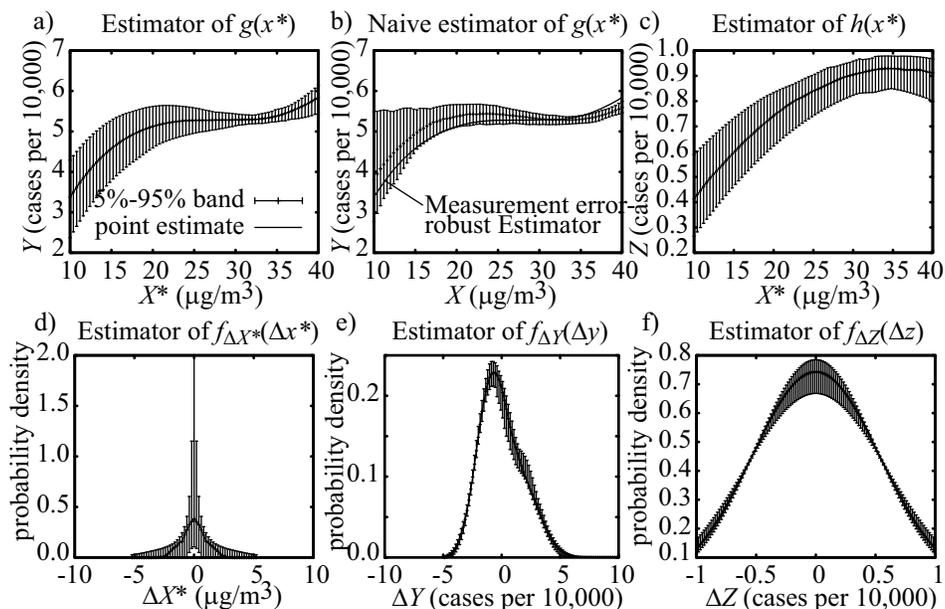


Figure 2: Application of the proposed estimator to an epidemiologic example (see text for a description of the variables and the estimated functions). In each plot, the estimator is shown as a solid line while the error bars indicate the pointwise 90% confidence bands. In b), the “naive” estimator is a nonparametric polynomial series least-square estimator that ignores the presence of measurement error. The estimator in a) is shown on the plot b) for comparison.

For comparison purposes, we also naively regress the dependent variables (Y or Z) on the mismeasured regressor X using a conventional least squares (thereby neglecting measurement error) with a polynomial specification with the same number of terms as our Berkson model. A first troubling observation from this exercise (see Figure 2b)) is that the naive estimate of $g(x^*)$ is not monotone, although in the region where it is unexpectedly decreasing, the confidence bands do not rule out a constant response. Second, it is perhaps counter-intuitive that the confidence bands for the naive estimator are sometimes larger than the corresponding bands for the measurement error-robust estimator. This is a consequence of the fact that correcting for Berkson errors amounts to an operation akin to convolution (rather than

deconvolution, as in classical measurement errors). Unlike deconvolution, convolution is a noise-reducing operation, effectively averaging observations of Y over a wide range of values of X to yield an estimate the expected value of Y given a specific value of X . This phenomenon is probably also responsible for the more reasonable (i.e. increasing) behavior of the response for the measurement error-robust estimate. Finally, the measurement error-robust regression function often lies at or beyond the 95% or 5% percentiles of the naive estimator distribution (see Figure 2b)). This implies that the level of any statistical test would be severely biased. For instance, the confidence bands of the naive estimator would reject our best estimate of $g(x^*)$ obtained with the measurement-error robust procedure.

In summary, this application example serves to illustrate that ignoring Berkson errors can be seriously misleading in nonlinear settings. Not only is the shape of the estimated response considerably affected, but statistical inferences based on a measurement error-blind method would be seriously biased. This application example also shows that our fully nonparametric and measurement error-robust method works well at sample sizes typically available in real data sets, without assuming the knowledge of the distribution of the measurement error.

A Identification Proof

Let $\mathcal{L}_1^b(\mathcal{D})$ with $\mathcal{D} \subset \mathbb{R}^{n_0}$ for some n_0 denote the set of all bounded functions in $\mathcal{L}_1(\mathcal{D})$ endowed with the usual \mathcal{L}_1 norm. Also, whenever we state an equality between functions in $\mathcal{L}_1^b(\mathcal{D})$, we mean that their difference is zero in the \mathcal{L}_1 norm.

We provide two proofs of Theorem 1. The first one, suggested by a referee, relies on the additional assumptions that (i) Z and X^* have the same dimension and (ii) h and its inverse are differentiable. Assumption (i) makes Assumption 3.3 unlikely to hold, but enables a somewhat direct application of Theorem 1 in Hu and Schennach (2008). The second proof relaxes those assumptions. It borrows some of the operator techniques from Hu and Schennach (2008), yet requires considerable changes in the approach — we focus here on the aspects of the proof that differ.

Proof Theorem 3.1 (simple special case). Let variables from Hu and Schennach (2008) be denoted by the corresponding uppercase letter with tildes and make the following assignments: $(\tilde{X}^*, \tilde{X}, \tilde{Y}, \tilde{Z}) = (h(X^*), Z, Y, X)$. We now verify the 5 assumptions of Theorem 1 in Hu and Schennach (2008).

To verify Assumption 1, we observe that the densities of $(\tilde{X}^*, \tilde{X}, \tilde{Y}, \tilde{Z})$ and (X^*, Z, Y, X) are related through: $f_{\tilde{X}^*, \tilde{X}, \tilde{Y}, \tilde{Z}}(\tilde{x}^*, \tilde{x}, \tilde{y}, \tilde{z}) = f_{X^*, Z, Y, X}(h^{-1}(\tilde{x}^*), \tilde{x}, \tilde{y}, \tilde{z}) |\partial h^{-1}(\tilde{x}^*) / \partial \tilde{x}^*|$ where the density $f_{X^*, Z, Y, X}$ exists by Assumption 3.1 and $h^{-1}(\tilde{x}^*)$ exists by Assumption 3.3. The Jacobian $\partial h^{-1}(\tilde{x}^*) / \partial \tilde{x}^*$ matrix is only defined if X^* and Z (and therefore \tilde{X}^*) have the same dimension and is finite and nonsingular under the assumption that h and its inverse are differentiable. A similar argument can be used for marginals and conditional distributions.

To verify Assumption 2, we note that our model can be written in terms of tilded variables

as:

$$\tilde{Y} = Y = g\left(h^{-1}\left(\tilde{X}^*\right)\right) + \Delta Y \quad (\text{A.1})$$

$$\tilde{Z} = X = h^{-1}\left(\tilde{X}^*\right) - \Delta X^* \quad (\text{A.2})$$

$$\tilde{X} = Z = \tilde{X}^* + \Delta Z. \quad (\text{A.3})$$

To verify Assumption 2 (i), we write

$$\begin{aligned} f_{\tilde{Y}|\tilde{X},\tilde{X}^*,\tilde{Z}}(\tilde{y}|\tilde{x},\tilde{x}^*,\tilde{z}) &= f_{Y|Z,X^*,X}(\tilde{y}|\tilde{x},h^{-1}(\tilde{x}^*),\tilde{z}) \\ &= f_{\Delta Y|\Delta Z,\Delta X^*,X}(\tilde{y}-g(h^{-1}(\tilde{x}^*))|\tilde{x}-\tilde{x}^*,h^{-1}(\tilde{x}^*)-\tilde{z},\tilde{z}) \\ &= f_{\Delta Y}(\tilde{y}-g(h^{-1}(\tilde{x}^*))) \\ &= f_{Y|\tilde{X}^*}(\tilde{y}|\tilde{x}^*) = f_{\tilde{Y}|\tilde{X}^*}(\tilde{y}|\tilde{x}^*) \end{aligned}$$

where we have used, in turn, (i) the equality $(\tilde{X}^*, \tilde{X}, \tilde{Y}, \tilde{Z}) = (h(X^*), Z, Y, X)$ and the fact that changes of variables in the conditioning variables do not introduce Jacobian terms, (ii) the fact that conditioning on Z, X^*, X is equivalent to conditioning on $\Delta Z, \Delta X^*, X$ (iii) Assumption 2.1, (iv) the relationship between ΔY and Y via (A.1) and (v) the equality $Y = \tilde{Y}$.

To verify Assumption 2 (ii), we similarly write

$$\begin{aligned} f_{\tilde{X}|\tilde{X}^*,\tilde{Z}}(\tilde{x}|\tilde{x}^*,\tilde{z}) &= f_{Z|X^*,X}(\tilde{x}|h^{-1}(\tilde{x}^*),\tilde{z}) \\ &= f_{\Delta Z|\Delta X^*,X}(\tilde{x}-\tilde{x}^*|h^{-1}(\tilde{x}^*)-\tilde{z},\tilde{z}) \\ &= f_{\Delta Z}(\tilde{x}-\tilde{x}^*) = f_{Z|\tilde{X}^*}(\tilde{x}|\tilde{x}^*) = f_{\tilde{X}|\tilde{X}^*}(\tilde{x}|\tilde{x}^*). \end{aligned}$$

Assumption 3 is implied by Assumptions 3.1, 2.1, 3.2, 3.3, 3.4 and Lemma A.1 below.

Assumption 4 requires that $f_{\tilde{Y}|\tilde{X}^*}(\tilde{y}|\tilde{x}_1^*) \neq f_{\tilde{Y}|\tilde{X}^*}(\tilde{y}|\tilde{x}_2^*)$ for $\tilde{x}_1^* \neq \tilde{x}_2^*$. This can be verified as follows:

$$\begin{aligned} f_{\tilde{Y}|\tilde{X}^*}(\tilde{y}|\tilde{x}_1^*) &= f_{\Delta Y|\tilde{X}^*}(\tilde{y}-g(h^{-1}(\tilde{x}_1^*))|\tilde{x}_1^*) \\ &= f_{\Delta Y}(\tilde{y}-g(h^{-1}(\tilde{x}_1^*))) \\ &\neq f_{\Delta Y}(\tilde{y}-g(h^{-1}(\tilde{x}_2^*))) = f_{\tilde{Y}|\tilde{X}^*}(\tilde{y}|\tilde{x}_2^*) \end{aligned}$$

by invoking (i) the definition of ΔY , (ii) independence of ΔY from X^* (and therefore \tilde{X}^*), (iii) the fact that $\tilde{x}_1^* \neq \tilde{x}_2^*$ implies $g(h^{-1}(\tilde{x}_1^*)) \neq g(h^{-1}(\tilde{x}_2^*))$ since $g(\cdot)$ and $h(\cdot)$ are one-to-one by Assumption 3.3 and so is $g(h^{-1}(\cdot))$.

Assumption 5 is trivially satisfied, by Equation (A.3).

Theorem 1 in Hu and Schennach (2008) then allows us to conclude that the joint distribution of $(h(X^*), X, Y, Z)$ is identified. However, in order to identify the distribution of (X^*, X, Y, Z) , we need to identify $h(\cdot)$. To this effect, we note that, conditional on $X = x$, the fluctuations in \tilde{X}^* are entirely caused by fluctuations in ΔX^* , by Equation (A.2). Moreover, ΔX^* is independent from X , hence,

$$f_{\tilde{X}^*|X}(\tilde{x}^*|x) = f_{\Delta X^*}(h^{-1}(\tilde{x}^*) - x) \left| \frac{\partial h^{-1}(\tilde{x}^*)}{\partial \tilde{x}^*} \right|, \quad (\text{A.4})$$

where the left-hand-side was previously identified and where the Jacobian term is well-defined by Assumptions 3.3 and the assumed differentiability of $h^{-1}(\tilde{x}^*)$. The Jacobian can be identified by integrating (A.4) with respect to x^* to yield: $\int f_{\tilde{X}^*|X}(\tilde{x}^*|x) dx = \left| \frac{\partial h^{-1}(\tilde{x}^*)}{\partial \tilde{x}^*} \right|$. By varying x while keeping \tilde{x}^* fixed in Equation (A.4), we can identify the density $f_{\Delta X^*}$ up to a shift of $h^{-1}(\tilde{x}^*)$. Assumption 2.2, pins down what the shift should be, so that $h^{-1}(\tilde{x}^*)$ is identified for any given \tilde{x}^* . Since $h(\cdot)$ is one-to-one by Assumption 3.3, $h^{-1}(\cdot)$ uniquely determines $h(\cdot)$. Hence, the joint distribution of (X^*, X, Y, Z) is identified. Finally, noting that $f_{Y|X^*}(y|x^*) = f_{\Delta Y}(y - g(x^*))$ (by Assumption 2.1), then establishes the identification of $g(x^*)$ with the help of Assumption 2.2. \square

Proof of Theorem 3.1 (general case). This proof borrows some of the operator techniques from Hu and Schennach (2008) and we focus here on the aspects of the proof that differ.

The definition of marginal and conditional densities in combination with Assumption 2.1 lead to the following sequence of equalities:

$$\begin{aligned}
f_{Y,Z|X}(y, z|x) &= \int f_{Y|X^*,Z,X}(y|x^*, z, x) f_{X^*,Z|X}(x^*, z|x) dx^* \\
&= \int f_{\Delta Y|X^*,\Delta Z,\Delta X^*}(y - g(x^*)|x^*, z - h(x^*), x^* - x) f_{X^*,Z|X}(x^*, z|x) dx^* \\
&= \int f_{\Delta Y}(y - g(x^*)) f_{X^*,Z|X}(x^*, z|x) dx^* \\
&= \int f_{\Delta Y}(y - g(x^*)) f_{Z|X^*,X}(z|x^*, x) f_{X^*|X}(x^*|x) dx^* \\
&= \int f_{\Delta Y}(y - g(x^*)) f_{\Delta Z|X^*,\Delta X^*}(z - h(x^*)|x^*, x^* - x) f_{\Delta X^*|X}(x^* - x|x) dx^* \\
&= \int f_{\Delta Y}(y - g(x^*)) f_{\Delta Z}(z - h(x^*)) f_{\Delta X^*}(x^* - x) dx^*
\end{aligned}$$

or, equivalently,

$$f_{Y,Z|X}(y, z|x) = \int f_{Z|X^*}(z|x^*) f_{Y|X^*}(y|x^*) f_{X^*|X}(x^*|x) dx^*. \quad (\text{A.5})$$

As in Hu and Schennach (2008), this integral equation can be written more conveniently as an operator equivalence relation

$$F_{y;Z|X} = F_{Z|X^*} D_{y;X^*} F_{X^*|X} \quad (\text{A.6})$$

by introducing the operators defined in Equation (3.5), which are acting on an arbitrary $r \in \mathcal{L}_1^b(\mathcal{X})$ (or $r \in \mathcal{L}_1^b(\mathcal{X}^*)$)

Similarly, one can show that

$$f_{Z|X}(z|x) = \int f_{Z|X^*}(z|x^*) f_{X^*|X}(x^*|x) dx^* \quad (\text{A.7})$$

and thus $F_{Z|X} = F_{Z|X^*} F_{X^*|X}$. By Assumptions 3.1, 2.1, 3.2, 3.3, 3.4 and Lemma A.1 below, we know that $F_{Z|X^*}$ admits an inverse on the range of $F_{Z|X^*}$ (and therefore the range of $F_{Z|X}$) and we can write

$$F_{X^*|X} = F_{Z|X^*}^{-1} F_{Z|X}. \quad (\text{A.8})$$

Substituting (A.8) into (A.6), we obtain:

$$F_{y;Z|X} = F_{Z|X^*} D_{y;X^*} F_{Z|X^*}^{-1} F_{Z|X}.$$

By Assumptions 3.1, 2.1, 3.2, 3.3, 3.4 and Lemma A.1 below again, $F_{Z|X}$ admits an inverse. Moreover, by Lemma 1 in Hu and Schennach (2008), the domain of $F_{Z|X}^{-1}$ is dense in $\mathcal{L}_1^b(\mathcal{Z})$ and we can then write

$$F_{y;Z|X} F_{Z|X}^{-1} = F_{Z|X^*} D_{y;X^*} F_{Z|X^*}^{-1}. \quad (\text{A.9})$$

Equation (A.9) states that the operator $F_{y;Z|X} F_{Z|X}^{-1}$ admits a spectral decomposition, where the eigenvalues are given by the $f_{Y|X^*}(y|x^*)$ for $x^* \in \mathcal{X}^*$ (for a fixed y) defining the operator $D_{y;X^*}$ while the eigenfunctions are the functions $f_{Z|X^*}(\cdot|x^*)$ for $x^* \in \mathcal{X}^*$ defining the kernel of the operator $F_{Z|X^*}$. As usual, the knowledge of a linear operator (e.g. $F_{Z|X}$) only determines the value of its kernel (e.g. $f_{Z|X}(z|x)$) everywhere except on a set of null Lebesgue measure. The resulting equivalence class exactly matches the usual equivalence class for probability densities with respect to the Lebesgue measure, so identifiability of the model is not affected.

The operator to be diagonalized is entirely defined in terms of observable densities while the decomposition provides the unobserved densities of interest. To ensure uniqueness of this decomposition, we employ four techniques. First, a powerful result from spectral analysis (Theorem XV 4.5 in Dunford and Schwartz (1971)) ensures uniqueness up to some normalizations. Second, the *a priori* arbitrary scale of the eigenfunctions is fixed by the requirement that densities must integrate to one. Third, to avoid any ambiguity in the definition of the eigenfunctions when degenerate eigenvalues are present, we use Assumption 3.3 and the fact that the eigenfunctions (which do not depend on y , unlike the eigenvalues $f_{y|x^*}(y|x^*)$) must be consistent across different values of the dependent variable y . These three steps are described in detail in Hu and Schennach (2008) and are not repeated here.

The fourth step (which differs from the approach taken in Hu and Schennach (2008)) is to rule out that the eigenvalues $f_{y;X^*}(y, x^*)$ and eigenfunctions $f_{Z|X^*}(\cdot|x^*)$ could be indexed by a different variable without affecting the operator $F_{y;Z|X} F_{Z|X}^{-1}$. (This issue is analogous to the nonunique ordering of the eigenvalues and eigenvectors in matrix diagonalization.) Suppose that the eigenfunctions can be indexed by another value, i.e., they are given by $f_{Z|\tilde{X}^*}(\cdot|\tilde{x}^*)$ where \tilde{x}^* is another variable related to x^* through $x^* = S(\tilde{x}^*)$ for some one-to-one function S .¹ Under this alternative indexing, all the assumptions of the original model must still hold with x^* replaced by \tilde{x}^* , so a relationship similar to (A.7) would still have to hold, for the same observed $f_{Z|X}(z|x)$:

$$f_{Z|X}(z|x) = \int f_{Z|\tilde{X}^*}(z|\tilde{x}^*) f_{\tilde{X}^*|X}(\tilde{x}^*|x) d\tilde{x}^*. \quad (\text{A.10})$$

or, in operator notation, $F_{Z|X} = F_{Z|\tilde{X}^*} F_{\tilde{X}^*|X}$.

In order for $f_{Z|\tilde{X}^*}(z|\tilde{x}^*)$ to be a valid alternative density, it must satisfy the same assumptions (and their implications) as $f_{Z|X^*}(z|x^*)$. In particular, the fact that $F_{Z|X^*}$ is invertible (established above via Lemma A.1) must also hold for $F_{Z|\tilde{X}^*}$. Hence, for any alternative $F_{Z|\tilde{X}^*}$, there is a unique corresponding $F_{\tilde{X}^*|X}$, given by $F_{\tilde{X}^*|X} = F_{Z|\tilde{X}^*}^{-1} F_{Z|X}$. We can find a

¹Note that $S(\cdot)$ is also measurable, for otherwise $X^* \equiv S(\tilde{X}^*)$ would not be a proper random variable.

more explicit expression for $f_{\tilde{X}^*|X}(\tilde{x}^*|x)$ as follows. First note that we trivially have that $f_{Z|\tilde{X}^*}(z|\tilde{x}^*) = f_{Z|X^*}(z|S(\tilde{x}^*))$ since $x^* = S(\tilde{x}^*)$ and S is one-to-one. By performing the change of variable $x^* = S(\tilde{x}^*)$ in (A.7), we obtain

$$f_{Z|X}(z|x) = \int f_{Z|X^*}(z|S(\tilde{x}^*)) f_{X^*|X}(S(\tilde{x}^*)|x) d\mu(\tilde{x}^*)$$

where the measure μ is defined, via $\mu(\mathcal{A}) = \lambda(S^{-1}(\mathcal{A}))$ for any measurable set \mathcal{A} , where λ denotes the Lebesgue measure and $S^{-1}(\mathcal{A}) \equiv \{\tilde{x}^* \in \mathcal{A} : S(\tilde{x}^*) = x^*\}$. From this we can conclude the equality between the two following measures

$$f_{\tilde{X}^*|X}(\tilde{x}^*|x) d\tilde{x}^* = f_{X^*|X}(S(\tilde{x}^*)|x) d\mu(\tilde{x}^*) \quad (\text{A.11})$$

by comparison with Equation (A.10) and the uniqueness of the measure $f_{\tilde{X}^*|X}(\tilde{x}^*|x) d\tilde{x}^*$ due to the injectivity of the $F_{Z|\tilde{X}^*}$ operator, shown in Lemma A.1 in the general case where the domain of $F_{Z|\tilde{X}^*}$ could include finite signed measures. We will now show that $f_{\tilde{X}^*|X}(\tilde{x}^*|x)$ necessarily violates Assumption 2.2 (with ΔX^* replaced by $\Delta\tilde{X}^* \equiv \tilde{X}^* - X$), unless $S(\cdot)$ is the identity function.²

Since $\Delta X^* = X^* - X$ with ΔX^* independent from X , we have $f_{X^*|X}(x^*|x) = f_{\Delta X^*}(x^* - x)$ and by a similar reasoning $f_{\tilde{X}^*|X}(\tilde{x}^*|x) = f_{\Delta\tilde{X}^*}(\tilde{x}^* - x)$ with $\Delta\tilde{X}^* \equiv \tilde{X}^* - X$. Equation (A.11) then becomes:

$$f_{\Delta\tilde{X}^*}(\tilde{x}^* - x) d\tilde{x}^* = f_{\Delta X^*}(S(\tilde{x}^*) - x) d\mu(\tilde{x}^*). \quad (\text{A.12})$$

Now, for a given x , consider Radom-Nikodym derivative of $f_{\Delta\tilde{X}^*}(\tilde{x}^* - x) d\tilde{x}^*$ with respect to the Lebesgue measure $d\tilde{x}^*$, which is, by definition, (almost everywhere) equal to $f_{\Delta\tilde{X}^*}(\tilde{x}^* - x)$, a bounded function by Assumption 3.1. By Equation (A.12), the existence of the Radom-Nikodym derivative of the left-hand side implies the existence of the same Radom-Nikodym derivative on the right-hand side and we can write:

$$f_{\Delta\tilde{X}^*}(\tilde{x}^* - x) = f_{\Delta X^*}(S(\tilde{x}^*) - x) \frac{d\mu(\tilde{x}^*)}{d\tilde{x}^*}. \quad (\text{A.13})$$

almost everywhere. Integrating both sides of the equation over all $x \in \mathcal{X}$, we obtain (after noting that points where the equality may fail have null measure and therefore do not contribute to the integral), $1 = 1 \frac{d\mu(\tilde{x}^*)}{d\tilde{x}^*}$, since densities integrate to 1, which implies that $d\mu(\tilde{x}^*)/d\tilde{x}^* = 1$, i.e. μ is also the Lebesgue measure. It follows from (A.13) that, almost everywhere

$$f_{\Delta\tilde{X}^*}(\tilde{x}^* - x) = f_{\Delta X^*}(S(\tilde{x}^*) - x).$$

In order for Assumption 2.2 to hold for both $\Delta\tilde{X}^*$ and ΔX^* , we must have that $f_{\Delta\tilde{X}^*}(\tilde{x}^* - x)$, when viewed as a function of \tilde{x}^* for any given x , is centered at $\tilde{x}^* = x$ and we must simultaneously have that $f_{\Delta X^*}(x^* - x) = f_{\Delta X^*}(S(\tilde{x}^*) - x)$, when viewed as a function of x^* for any given x , is centered at $x^* = x$, i.e. $S(\tilde{x}^*) = x$. The two statements are only compatible if $\tilde{x}^* = S(\tilde{x}^*)$. Thus, there cannot exist two distinct but observationally equivalent parametrization of the eigenvalues/eigenfunctions.

²Some of the steps below were inspired by comments from an anonymous referee.

Hence we have shown, through Equation (A.9), that the unobserved functions $f_{Y|X^*}(y|x^*)$ and $f_{Z|X^*}(\cdot|x^*)$ are uniquely determined (up to an equivalence class of functions differing at most on a set of null Lebesgue measure) by the observed function $f_{Y,Z|X}(y,z|x)$. Next, Equation (A.8) implies that $f_{X^*|X}(x^*|x)$ is uniquely determined as well.

Once $f_{Y|X^*}(y|x^*)$ and $f_{Z|X^*}(z|x^*)$ are known, the functions $g(x^*)$ and $h(x^*)$ can be identified by exploiting the centering restrictions on ΔY , ΔX^* and ΔZ , e.g. $g(x^*) = \int y f_{Y|X^*}(y|x^*) dy$ if ΔY is assumed to have zero mean. Next, $f_{\Delta Y}(\Delta y)$ can be straightforwardly identified, e.g. $f_{\Delta Y}(\Delta y) = f_{Y|X^*}(g(x^*) + \Delta y|x^*)$ for any $x^* \in \mathcal{X}^*$. Similar arguments yield $h(x^*)$ and $f_{\Delta Z}(\Delta z)$ from $f_{Z|X^*}(z|x^*)$ as well as $f_{\Delta X^*}(\Delta x^*)$ from $f_{X^*|X}(x^*|x)$. It follows that Equation (3.1) has a unique solution. The second conclusion of the Theorem then follows from the fact that both $f_{Y,Z|X}(y,z|x)$ and $f_X(x)$ are uniquely determined (except perhaps on a set of null Lebesgue measure) from $f_{Y,Z,X}(y,z,x)$. \square

The following Lemma is closely related to Proposition 2.4 in d'Haultfoeuille (2011). It is different in terms of the spaces the operators can act on and more general in terms of the possible dimensionalities of the random variables involved.

Lemma A.1. *Let X, X^* and Z be generated by Equations (2.2)-(2.3). Let $\mathcal{S}(\mathcal{T})$ be the set of finite signed measures on a given set $\mathcal{T} = \mathcal{X}, \mathcal{X}^*$ or \mathcal{Z} . (and note that $\mathcal{S}(\mathcal{T})$ includes $\mathcal{L}_1^b(\mathcal{T})$ as a special case, in the sense that for any function in $r \in \mathcal{L}_1^b(\mathcal{T})$, there is a corresponding measure $R \in \mathcal{S}(\mathcal{T})$ whose Radom-Nikodym derivative with respect to the Lebesgue measure is r). Under Assumptions 2.1, 3.1, 3.2, 3.3 and 3.4, the operators $F_{X^*|X} : \mathcal{S}(\mathcal{X}) \mapsto \mathcal{L}_1^b(\mathcal{X}^*)$, $F_{Z|X^*} : \mathcal{S}(\mathcal{X}^*) \mapsto \mathcal{L}_1^b(\mathcal{Z})$ and $F_{Z|X} : \mathcal{S}(\mathcal{X}) \mapsto \mathcal{L}_1^b(\mathcal{Z})$, defined in (3.5), are injective mappings.*

Proof. First, one can verify that $R \in \mathcal{S}(\mathcal{X})$ implies that $F_{X^*|X}R \in \mathcal{L}_1^b(\mathcal{X}^*)$ and similarly for $F_{Z|X^*}$ and $F_{Z|X}$, since the (conditional) densities involving variables X^*, X and Z are bounded by Assumption 3.1 and are absolutely integrable. We now verify injectivity of $F_{Z|X^*}$.

By Assumptions 2.1, 3.1 and Equation (2.3), we have, for any $R \in \mathcal{S}(\mathcal{X}^*)$,

$$[F_{Z|X^*}R](z) = \int f_{Z|X^*}(z|x^*) dR(x^*) = \int f_{\Delta Z}(z - h(x^*)) dR(x^*).$$

Next, let \tilde{R} denote the signed measure assigning, to any measurable set $\mathcal{A} \subseteq \mathbb{R}^{n_z}$, the value $\tilde{R}(\mathcal{A}) = \int 1(h(x^*) \in \mathcal{A}) dR(x^*)$ and note that \tilde{R} is a finite signed measure since $R(x^*)$ is. Then, we can express $F_{Z|X^*}R$ as

$$[F_{Z|X^*}R](z) = \int f_{\Delta Z}(z - \tilde{x}^*) d\tilde{R}(\tilde{x}^*), \quad (\text{A.14})$$

i.e. a convolution between the probability measure of ΔZ (represented by its Lebesgue density) and the signed measure \tilde{R} (see Chapter 5 in Bhattacharya and Rao (2010)). By the convolution Theorem for signed measures (Theorem 5.1(iii) in Bhattacharya and Rao (2010)), one can convert the convolution (A.14) into a product of Fourier transforms:³

$$\sigma(\zeta) = \phi_{\Delta Z}(\zeta) \rho(\zeta)$$

³Note that the Fourier transforms involved are all continuous functions because the original functions (or measures) are absolutely integrable (or finite), hence ‘‘almost everywhere’’ qualifications do not apply to them.

where $\sigma(\zeta) \equiv \int [F_{Z|X^*} R](z) e^{i\zeta z} dz$, $\phi_{\Delta Z}(\zeta) \equiv E[e^{i\zeta Z}]$ and $\rho(\zeta) \equiv \int e^{i\zeta z} d\tilde{R}(z)$. Since $\phi_{\Delta Z}(\zeta)$, the characteristic function of ΔZ , is nonvanishing by Assumption 3.2, we can isolate $\rho(\zeta)$ as

$$\rho(\zeta) = \sigma(\zeta) / \phi_{\Delta Z}(\zeta).$$

Since there is a one-to-one mapping between finite signed measures and their Fourier transforms (by Theorem 5.1(i) in Bhattacharya and Rao (2010)), \tilde{R} can be recovered as the unique signed measure whose Fourier transform is $\rho(\zeta)$. We now show that the signed measure \tilde{R} uniquely determines the measure R .

Let $\mathcal{A}_{\mathcal{B}} = \cup_{x^* \in \mathcal{B}} \{h(x^*)\}$ for any measurable $\mathcal{B} \subseteq \mathbb{R}^{n_x}$ and note that $\mathcal{A}_{\mathcal{B}}$ is also measurable since h is continuous by Assumption 3.4. Then, observe that, by Assumption 3.3, $h(x^*) \in \mathcal{A}_{\mathcal{B}}$ iff $x^* \in \mathcal{B}$ and we have:

$$\tilde{R}(\mathcal{A}_{\mathcal{B}}) = \int 1(h(x^*) \in \mathcal{A}_{\mathcal{B}}) dR(x^*) = \int 1(x^* \in \mathcal{B}) dR(x^*).$$

Since \mathcal{B} is arbitrary, the knowledge of $\tilde{R}(\mathcal{A}_{\mathcal{B}})$ uniquely determines the value assigned to any measurable set by the signed measure R .

Injectivity of $F_{X^*|X}$ is a special case of the above derivation (with Z, X^* replaced by X^*, X), in which h is the identity function. Finally, injectivity of $F_{Z|X}$ is implied by the injectivity of $F_{Z|X^*}$ and $F_{X^*|X}$, since $F_{Z|X} = F_{Z|X^*} F_{X^*|X}$ by Assumption 2.1 and Equations (2.2)-(2.3). \square

B Consistency Proof

The proof of Theorem 4.1 in the main text relies on the following 4 simple Lemmas.

Lemma B.1. *The set \mathcal{S} is compact in the norm $\|\cdot\|$ (see Definition 4.1).*

Proof. \mathcal{S} is closed by construction, hence compactness follows from showing that \mathcal{S} can be covered by a finite number of $\|\cdot\|$ -balls of any radius $\varepsilon > 0$. For a given $\varepsilon > 0$, let $\bar{v} > 0$ be such that $\varepsilon \geq \int_{\bar{v}}^{\infty} f'_+(v) dv$. Such a \bar{v} can always be found because $\int_{-\infty}^{\infty} f'_+(v) dv < \infty$ by assumption. Then, let n be the smallest integer such that $n \geq 2\bar{v} f'_+(0) / \varepsilon$ and define the partition $\mathcal{P}_{\varepsilon} =]-\infty, -\bar{v}] , [-\bar{v}, -\bar{v}(n-2)/n] , \dots , [\bar{v}(n-2)/n, \bar{v}] , [\bar{v}, \infty[$. This partition is such that any function $f \in \mathcal{S}$ cannot vary by more than ε over each interval of the partition: For the two infinite end intervals, this follows from (without loss of generality, consider $v_2 > v_1 > \bar{v}$)

$$|f(\bar{v}) - f(v)| \leq \int_{v_1}^{v_2} |f'(v)| dv \leq \int_{v_1}^{v_2} f'_+(v) dv \leq \int_{\bar{v}}^{\infty} f'_+(v) dv \leq \varepsilon$$

while for each of the n finite intervals, this follows from (without loss of generality, consider $v_2 > v_1$ and $v_1, v_2 \in [\bar{v}(n-2)/n, \bar{v}]$)

$$|f(v_2) - f(v_1)| \leq \int_{v_1}^{v_2} |f'(v)| dv \leq \int_{v_1}^{v_2} f'_+(0) dv = (v_2 - v_1) f'_+(0) \leq \varepsilon$$

since $f'_+(v)$ is decreasing in $|v|$ and since $(v_2 - v_1) \leq \bar{v} - \bar{v}(n-2)/n = 2\bar{v}/n = 2\bar{v}/(2\bar{v}f'_+(0)/\varepsilon) = \varepsilon/(f'_+(0))$.

Next, let m be the smallest integer such that $m \geq 2B/\varepsilon$ and define the finite set

$$\mathcal{R}_\varepsilon = \{-B, -B(m-2)/m, \dots, B(m-2)/m, B\}$$

and note that consecutive elements are no further than ε apart. It follows that any function $f \in \mathcal{S}$ can be approximated with an error no larger than ε by a function $\tilde{f} : \mathbb{R} \mapsto \mathcal{R}_\varepsilon$ that is piecewise constant on the intervals of the partition \mathcal{P}_ε . There are m^{n+2} such functions, which is a finite number for any $\varepsilon > 0$ and hence \mathcal{F} is compact. \square

Lemma B.2. *The sets \mathcal{F} and \mathcal{G} are compact in the norms $\|\cdot\|$ and $\|\cdot\|_\omega$, respectively.*

Proof. Since $\mathcal{F} \subseteq \mathcal{S}$ and \mathcal{S} is compact by Lemma B.1 while \mathcal{F} is closed, it follows that \mathcal{F} is compact as well. Let $\mathcal{S}_\omega = \{g : \|g\|_\omega \leq B\}$ and note that \mathcal{G} is closed and $\mathcal{G} \subseteq \mathcal{S}_\omega$. Since the mapping $f \mapsto \omega f$ from $(\mathcal{S}, \|\cdot\|)$ to $(\mathcal{S}_\omega, \|\cdot\|_\omega)$ is an isometry, \mathcal{S}_ω is also compact. It follows that \mathcal{G} is compact. \square

Lemma B.3. *If $f_{\Delta Y}, f_{\Delta Z}, f_{\Delta X^*} \in \mathcal{F}$ and $g, h \in \mathcal{G}$, then, for any $b > 0$ and any $x, y, z \in \mathbb{R}$, we have*

$$\int f_{\Delta Y}(y - g(x^*)) f_{\Delta Z}(z - h(x^*)) f_{\Delta X^*}(x - x^*) dx^* \geq f_-(x, y, z) > 0,$$

where

$$f_-(x, y, z) \equiv 2bf_-(b) f_-(|y| + (g_+(|x| + b))) f_-(|z| + (g_+(|x| + b))).$$

Proof. We have, for any $b > 0$,

$$\begin{aligned} & \int f_{\Delta Y}(y - g(x^*)) f_{\Delta Z}(z - h(x^*)) f_{\Delta X^*}(x - x^*) dx^* \\ & \geq \int f_-(y - g(x^*)) f_-(z - h(x^*)) f_-(x - x^*) dx^* \\ & \geq \int_{x-b}^{x+b} f_-(y - g(x^*)) f_-(z - h(x^*)) f_-(x - x^*) dx^* \\ & \geq 2bf_-(b) \inf_{x^* \in [x-b, x+b]} f_-(y - g(x^*)) f_-(z - h(x^*)) \\ & \geq 2bf_-(b) f_-(|y| + (g_+(|x| + b))) f_-(|z| + (g_+(|x| + b))) \equiv f_-(x, y, z) \end{aligned}$$

where we have used the fact that all densities in \mathcal{F} are bounded below by $f_-(v)$, that the integrand is positive, that $f_-(v)$ is symmetric and decreasing in $|v|$ and that $|y - g(x^*)| \leq |y| + |g(x^*)| \leq |y| + |g_+(x^*)| \leq |y| + |g_+(|x| + b)|$ for $x^* \in [x - b, x + b]$ (since $g_+(x^*)$ bounds $|g(x^*)|$ and is increasing) and similarly for $z - h(x^*)$. Finally $f_-(x, y, z) > 0$ since $f_-(v)$ is strictly positive for $v \in \mathbb{R}$. \square

Lemma B.4. *(This restates Lemma A1 in Newey and Powell (2003) for convenient reference). Suppose (i) $Q(\theta)$ has a unique maximum on Θ at θ^* , (ii) Θ is compact, (iii) $\hat{Q}(\theta)$ is continuous, (iv) $\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{p} 0$, (v) $Q(\theta)$ is continuous (vi) $\hat{\Theta}_n$ are compact subsets of Θ such that for any $\theta \in \Theta$, there exists a sequence $\{\tilde{\theta}_n\}$ with $\tilde{\theta}_n \in \hat{\Theta}_n$ such that $\tilde{\theta}_n \xrightarrow{p} \theta$. Then $\hat{\theta} = \arg \min_{\theta \in \hat{\Theta}} \hat{Q}(\theta) \xrightarrow{p} \theta^*$.*

Proof of Theorem 4.1. We verify the conditions of Lemma B.4 for an objective function $Q(\theta) = E[\ln f_{Y,Z|X}(Y, Z|X)]$ with $f_{Y,Z|X}(y, z|x)$ given by Equation (3.1) and $\hat{Q}(\theta)$ set to (4.3) with $\theta \equiv (g, h, f_{\Delta X^*}, f_{\Delta Y}, f_{\Delta Z}) \in \Theta \equiv \mathcal{G} \times \mathcal{G} \times \mathcal{F} \times \mathcal{F} \times \mathcal{F}$. The set Θ is endowed with the norm

$$\|\theta\| = \max\{\|g\|_\omega, \|h\|_\omega, \|f_{\Delta X^*}\|, \|f_{\Delta Y}\|, \|f_{\Delta Z}\|\}. \quad (\text{B.1})$$

The sets $\hat{\Theta}_n$ are the intersection of Θ with the span of the series (4.1) for $m = g, h$ and (4.2) for $V = \Delta X^*, \Delta Y, \Delta Z$ truncated at progressively larger n -dependent values of K_V (for $V = g, h, \Delta X^*, \Delta Y, \Delta Z$).

(i) $Q(\theta)$ is uniquely maximized at θ^* . By Theorem 3.1, for a given density of the data $f_{Y,Z|X}^*(y, z|x)$, there exists a unique solution to Equation (3.1), denoted $\theta^* \equiv (g^*, h^*, f_{\Delta X^*}^*, f_{\Delta Y}^*, f_{\Delta Z}^*)$. The expected likelihood $Q(\theta)$ is therefore uniquely maximized at θ^* , by the usual Jensen's inequality argument:

$$\begin{aligned} Q(\theta) - Q(\theta^*) &= E\left[\ln \frac{f(y, z|x)}{f^*(y, z|x)}\right] \\ &= \int \int \int \ln \frac{f_{YZ|X}(y, z|x)}{f_{YZ|X}^*(y, z|x)} f_{YZ|X}^*(y, z|x) dy dz f_X(x) dx \\ &\leq \int \ln \left(\int \int \frac{f_{YZ|X}(y, z|x)}{f_{YZ|X}^*(y, z|x)} f_{YZ|X}^*(y, z|x) dy dz \right) f_X(x) dx \\ &= \int \ln(1) f_X(x) dx = 0 \end{aligned}$$

with equality only if $f_{YZ|X}(y, z|x) / f_{YZ|X}^*(y, z|x)$ is almost everywhere constant, i.e. $f_{YZ|X}(y, z|x) = f_{YZ|X}^*(y, z|x)$. Since all densities are assumed differentiable by Definition 4.1, the equality actually holds everywhere.

(ii) Θ is compact. As \mathcal{F} and \mathcal{G} are compact by Lemma B.2, it follows that $\Theta \equiv \mathcal{G} \times \mathcal{G} \times \mathcal{F} \times \mathcal{F} \times \mathcal{F}$ is also compact in the norm (B.1) (finitely repeated cartesian products of finite coverings of \mathcal{F} and \mathcal{G} cover Θ as well).

(iii) $\hat{Q}(\theta)$ is continuous in θ . Note that $\hat{Q}(\theta) = n^{-1} \sum_{i=1}^n Q(\theta, X_i, Y_i, Z_i)$ where

$$Q(\theta, x, y, z) \equiv \ln \int f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - g(x^*)) f_{\Delta X^*}(x^* - x) dx^*,$$

so it is sufficient to show that $Q(\theta, x, y, z)$ is continuous in θ at all $x, y, z \in \mathbb{R}$. We verify continuity with respect to $f_{\Delta Z}, f_{\Delta Y}, f_{\Delta X^*}, g, h$ in turn, which will imply continuity jointly in all parameters, by the triangle inequality. Consider the change in $Q(\theta, x, y, z)$, denoted $\delta Q(\theta, x, y, z)$, due to a small change in $f_{\Delta Z}(\Delta z)$, denoted $\delta f_{\Delta Z}(\Delta z)$ and satisfying $\|\delta f_{\Delta Z}(\Delta z)\| \leq \varepsilon$:

$$\begin{aligned} &\delta Q(\theta, x, y, z) \\ &= \ln \int f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - g(x^*)) f_{\Delta X^*}(x^* - x) dx^* \\ &\quad - \ln \int f_{\Delta Z}(z - h(x^*)) + \delta f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - g(x^*)) f_{\Delta X^*}(x^* - x) dx^* \\ &= \frac{\int \delta f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - g(x^*)) f_{\Delta X^*}(x^* - x) dx^*}{\int f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - g(x^*)) f_{\Delta X^*}(x^* - x) dx^*} \end{aligned}$$

where $\dot{f}_{\Delta Z}$ denotes a mean value on the segment joining $f_{\Delta Z}$ and $f_{\Delta Z} + \delta f_{\Delta Z}$. Since the set \mathcal{F} is convex, $\dot{f}_{\Delta Z}$ shares the same inequality constraints as $f_{\Delta Z}$ and $f_{\Delta Z} + \delta f_{\Delta Z}$, for instance, $\dot{f}_{\Delta Z}(\Delta z) \geq f_-(\Delta z)$. Moreover, all densities in \mathcal{F} are assumed bounded by some constant B and $\|\delta f_{\Delta Z}(\Delta z)\| \leq \varepsilon$ implies $|\delta f_{\Delta Z}(z - h(x^*))| \leq \varepsilon$ while the denominator can be bounded below via Lemma B.3. We can then write:

$$\begin{aligned} |\delta Q(\theta, x, y, z)| &\leq \frac{\int |\delta f_{\Delta Z}(z - h(x^*))| f_{\Delta Y}(y - g(x^*)) f_{\Delta X^*}(x^* - x) dx^*}{\int f_-(z - h(x^*)) f_-(y - g(x^*)) f_-(x^* - x) dx^*} \\ &\leq \frac{\int \varepsilon B f_{\Delta X^*}(x^* - x) dx^*}{\int f_-(z - h(x^*)) f_-(y - g(x^*)) f_-(x^* - x) dx^*} \\ &\leq \frac{\varepsilon B \int f_{\Delta X^*}(x^* - x) dx^*}{f_-(x, y, z)} = \frac{\varepsilon B}{f_-(x, y, z)} \end{aligned} \quad (\text{B.2})$$

where $f_-(x, y, z) > 0$ by Lemma B.3.

We can bound the effect of changes in $f_{\Delta Y}$ on $Q(\theta, x, y, z)$ in an entirely analogous way. We then focus on the effect of changes in $f_{\Delta X^*}$, which requires a slightly different approach (because $f_{\Delta Z}(z - h(x^*))$ and $f_{\Delta Y}(y - g(x^*))$ are not necessarily integrable over x^*). As before, we have

$$\begin{aligned} \delta Q(\theta, x, y, z) &= \frac{\int f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - g(x^*)) \delta f_{\Delta X^*}(x^* - x) dx^*}{\int f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - g(x^*)) \dot{f}_{\Delta X^*}(x^* - x) dx^*} \\ &\leq \frac{B^2 \int \delta f_{\Delta X^*}(x^* - x) dx^*}{f_-(x, y, z)}. \end{aligned}$$

Now, since both $f_{\Delta X^*}$ and $f_{\Delta X^*} + \delta f_{\Delta X^*}$ are in \mathcal{F} and therefore bounded by an integrable function f_+ , we have that $|\delta f_{\Delta X^*}(v)|$ must be bounded by the integrable function $2f_+(v)$. By Lebesgue's Dominated Convergence Theorem, it follows that, for any sequence $\{\delta f_{\Delta X^*, m}\}$ with $\|\delta f_{\Delta X^*, m}\| \rightarrow 0$ (and therefore $\delta f_{\Delta X^*, m}(v) \rightarrow 0$ at each $v \in \mathbb{R}$), $\lim_{m \rightarrow \infty} \int \delta f_{\Delta X^*, m}(x^* - x) dx^* = \int \lim_{m \rightarrow \infty} \delta f_{\Delta X^*, m}(x^* - x) dx^* = 0$. Hence $Q(\theta, x, y, z)$ is continuous in $f_{\Delta X^*}$.

We now bound the effect, on $Q(\theta, x, y, z)$, of changes in $g(x^*)$, denoted by $\delta g(x^*)$ and satisfying $\|\delta g\|_\omega \leq \varepsilon$. We have

$$\delta Q(\theta, x, y, z) = \frac{\int f_{\Delta Z}(z - h(x^*)) f'_{\Delta Y}(y - \dot{g}(x^*)) \delta g(x^*) f_{\Delta X^*}(x^* - x) dx^*}{\int f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - \dot{g}(x^*)) f_{\Delta X^*}(x^* - x) dx^*}$$

where $\dot{g}(x^*)$ is a mean value on the segment jointing $g(x^*)$ and $g(x^*) + \delta g(x^*)$ which satisfies the same constraints as any $g(x^*)$, by convexity. As before, we have

$$\begin{aligned} |\delta Q(\theta, x, y, z)| &\leq \frac{\int f_{\Delta Z}(z - h(x^*)) |f'_{\Delta Y}(y - \dot{g}(x^*))| |\delta g(x^*)| f_{\Delta X^*}(x^* - x) dx^*}{\int f_-(z - h(x^*)) f_-(y - \dot{g}(x^*)) f_-(x^* - x) dx^*} \\ &\leq \frac{\int B |f'_{\Delta Y}(y - \dot{g}(x^*))| |\delta g(x^*)| f_{\Delta X^*}(x^* - x) dx^*}{f_-(x, y, z)} \\ &\leq \frac{BB_2 \varepsilon \int (\omega(x^*))^{-1} f_+(x^* - x) dx^*}{f_-(x, y, z)} = \varepsilon \frac{BB_2 \int (\omega(x^*))^{-1} f_+(x^* - x) dx^*}{f_-(x, y, z)} \end{aligned}$$

where we have used Assumption 4.4 and the facts that $\|\delta g\| \leq \varepsilon \implies |\delta g(x^*)| \leq \varepsilon/\omega(x^*)$ as well as $f_{\Delta X^*}(v) \leq f_+(v)$ and $|f'_{\Delta Y}(v)| \leq f'_+(v) \leq B_2$ for some $B_2 < \infty$. Since $\int (\omega(x^*))^{-1} f_+(x^* - x) dx^*$ is finite at each x by Assumption 4.4 and since $f_-(x, y, z) > 0$ by Lemma B.3, it follows that $Q(\theta, x, y, z)$ is continuous in g . A similar reasoning can be used to show continuity in h .

(iv) To show $\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \xrightarrow{p} 0$, we verify the conditions of Lemma 2.4 in Newey and McFadden (1994). Θ was already shown to be compact and $\hat{Q}(\theta)$ was already shown to be a sample average of $Q(\theta, X_i, Y_i, Z_i)$, where $Q(\theta, x, y, z)$ is a function everywhere continuous in θ . There only remains to show that $Q(\theta, x, y, z)$ can be bounded by a positive function $\bar{Q}(x, y, z)$ with $E[\bar{Q}(X, Y, Z)] < \infty$ and that does not depend on θ :

$$\begin{aligned} |Q(\theta, x, y, z)| &= \left| \ln \int f_{\Delta Z}(z - h(x^*)) f_{\Delta Y}(y - g(x^*)) f_{\Delta X^*}(x^* - x) dx^* \right| \\ &\leq \max \left\{ \left| \ln \int BB f_{\Delta X^*}(x^* - x) dx^* \right|, \right. \\ &\quad \left. \left| \ln \int f_-(z - h(x^*)) f_-(y - g(x^*)) f_-(x^* - x) dx^* \right| \right\} \\ &\leq \max \{ |\ln B^2|, |\ln f_-(x, y, z)| \} \equiv \bar{Q}(x, y, z) \end{aligned}$$

where we have used Lemma B.3. $E[\bar{Q}(X, Y, Z)] < \infty$ since B is finite and nonzero and $E[|\ln f_-(X, Y, Z)|] < \infty$ by Assumption 4.5.

(v) The fact that $Q(\theta)$ is continuous in θ follows from the fact that $\hat{Q}(\theta)$ is continuous (see (iii)) and converges uniformly (see (iv)).

(vi) The fact there exists, for any $\theta \in \Theta$, a sequence $\{\tilde{\theta}_n\}$ with $\tilde{\theta}_n$ in the compact set $\hat{\Theta}_n$ such that $\tilde{\theta}_n \xrightarrow{p} \theta$, follows directly from Assumption 4.3. Indeed, if the set of functions representable as series (4.2) is dense in \mathcal{F} (in the norm $\|\cdot\|$), there exists, for any $f \in \mathcal{F}$, a deterministic decreasing sequence $\varepsilon_k \rightarrow 0$ such that $\|f_k - f\| \leq \varepsilon_k$, where f_k denotes a partial sum of k terms of the series (4.2). For a given $\varepsilon > 0$, There exists a K guaranteeing that $\varepsilon_k \leq \varepsilon$ for all $k \geq K$. Since $K_V \xrightarrow{p} \infty$ for $V = \Delta X^*, \Delta Y, \Delta Z$, it follows that all K_V will eventually exceed that K with probability approach one. Hence, $\|f_{K_V} - f\| \xrightarrow{p} 0$ for $V = \Delta X^*, \Delta Y, \Delta Z$. A similar reasoning holds for g and h with the series (4.1). \square

Proof of Lemma 4.1 in the main text. If $\|f\| \equiv \sup_{v \in \mathbb{R}} |f(v)|$ takes a certain positive value, there must be at least one point v_0 such $|f(v_0)| = \|f\| - \varepsilon \equiv a$ for any $\varepsilon > 0$ sufficiently small. Away from that point, the function $f(v)$ cannot decrease faster than $f'_+(0)$, by the Lipschitz constraint on functions in \mathcal{F} . Hence the smallest possible value of $\int |f(v)|^2 dv$ will be reached for a triangular function $f_0(v) = \max\{a - b|v - v_0|, 0\}$. Without loss of generality, $v_0 = 0$ and we then have

$$\int |f(v)|^2 dv \geq \int_{-a/f'_+(0)}^{a/f'_+(0)} (a - f'_+(0)|v|)^2 dv = \frac{2}{3f'_+(0)} a^3$$

This is true for a arbitrarily close to $\|f\|$, therefore, $\|f\| \leq (3/2) f'_+(0) (\int |f(v)|^2 dv)^{1/3}$ and it follows that $\int |f_n(v)|^2 dv \rightarrow 0 \implies \|f_n\| \rightarrow 0$. \square

C Additional simulations

A second simulation example illustrates performances of our estimator in a different set of unfavorable circumstances. First, the distribution of the measurement error ΔX^* is a mixture of two normals with standard deviation $1/4$ centered at $-1/5$ and $2/5$ and with respective weights $2/3$ and $1/3$ (so that ΔX^* has zero mean). This is an asymmetric and nearly bimodal measurement error distribution that is quite challenging to tackle, because a very flexible sieve is required to model it. In fact, such complex cases have rarely been considered in benchmarking other methods to correct for measurement error bias in most classical error models. Second, X is distributed according to a symmetric triangular distribution on $[-1, 1]$. The standard deviation of this distribution is only 0.41 , which is not much bigger than the standard deviation of the measurement error distribution (0.38), thus making this estimation problem exceedingly difficult: The observable X contains almost as much noise as there is signal. Third, the distribution of ΔY is a Student t distribution with 4 degrees of freedom, divided by 4, which is heavy-tailed distribution that often leads to large estimator variability. Fourth, the distribution of ΔZ is a normal with mean 0 and standard deviation of 0.25 , which is of the same order of magnitude as the standard deviation of X (0.41), thus making Z a relatively uninformative instrument and making the estimation problem more difficult. Fifth, a commonly used logistic regression function is used to generate the data:

$$g(X^*) = (1 + \exp(-4X^*))^{-1}, \quad (\text{C.1})$$

which is highly nonlinear over the range of values of X^* that are heavily sampled. Finally, the instrument equation has an exponential specification:

$$h(X^*) = \frac{1}{2} \exp(X^*), \quad (\text{C.2})$$

which is strictly convex and therefore tends to exacerbate the bias in many nonparametric estimators. The estimation methodology is as described in Section 5, except that the optimal numbers of parameters were found to be $f_{\Delta X^*} : 6$; $f_{\Delta Y} : 5$; $f_{\Delta Z} : 4$; $g : 6$; $h : 3$.

Figure 3 summarizes the result of these simulations, where a naive nonparametric series least-squares estimator ignoring measurement error (i.e. least-square regressions of Y on X and of Z on X) with the same number of sieve terms is also shown for comparison. The reliability of the method can be appreciated by noting how closely the median of the replicated measurement-error-robust estimates matches the true model. In comparison, the median of the naive estimator is so significantly biased that any type of hypothesis test based on it would exhibit completely misleading confidence levels: The true model curves (for g and h) often lies beyond the 95% or 5% percentiles of the naive estimator distribution. Overall, the proposed measurement-error-robust estimator exhibits low variability and low bias at the reasonable sample size of 500. It has some difficulty estimating the intricate density of ΔX^* , as expected. Nevertheless, the estimates of the function of main interest ($g(\cdot)$) are quite reliable, despite the number of “difficult” features we have incorporated into this test case.

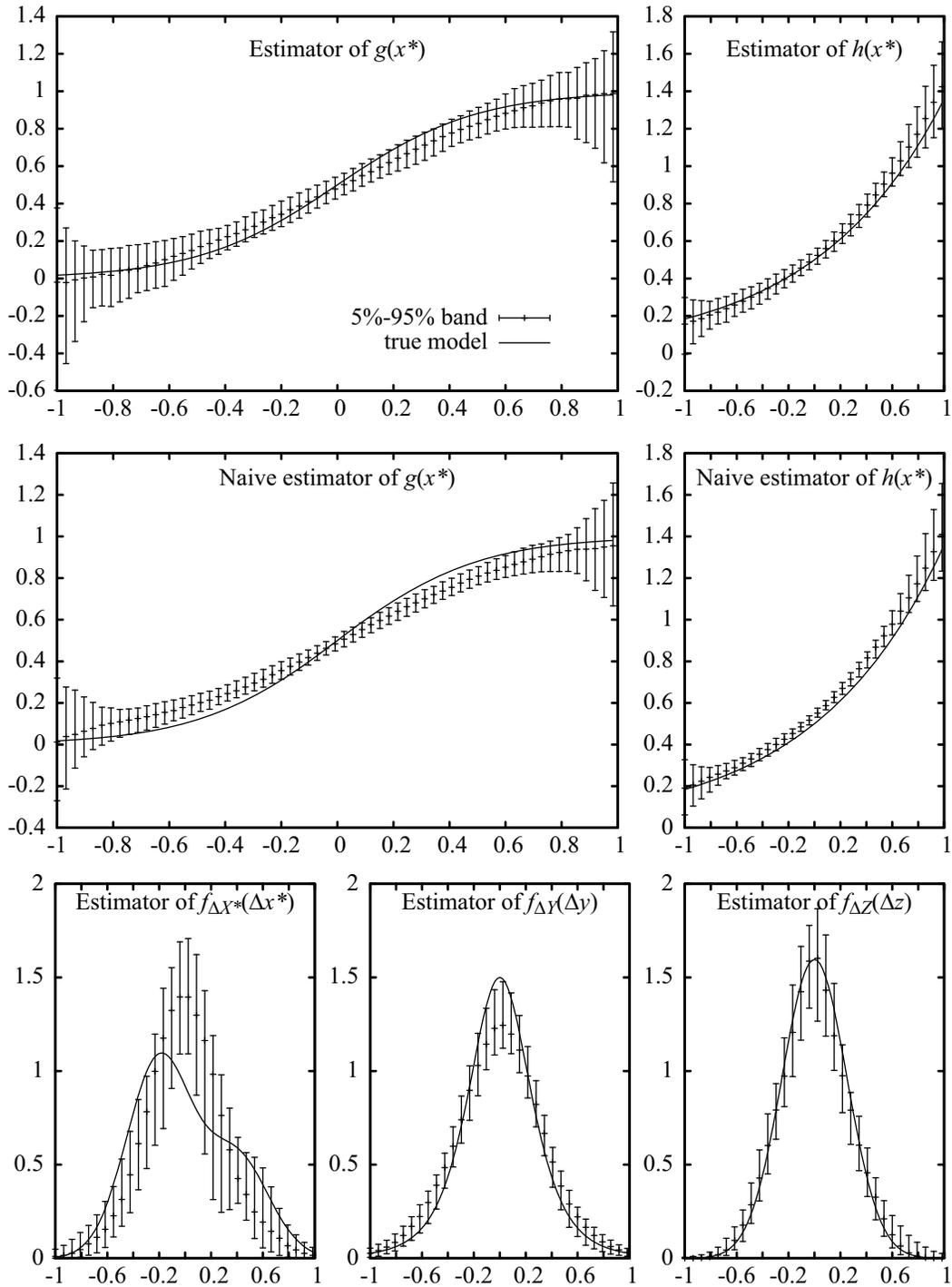


Figure 3: Simulation study of the practical performance of the proposed measurement-error-robust estimator in comparison with a “naive” nonparametric polynomial series least-square estimator that ignores the presence of measurement error. In each plot, the pointwise 90% confidence band of the estimator simulated over 100 replications is shown as error bars.

D Miscellaneous extensions

It is possible to relax Assumption 2.1, by only requiring the three quantities ΔY , ΔZ and $(X, \Delta X^*)$ to be mutually independent, provided that the centering restriction on ΔX^* imposes that the median of ΔX^* given X is unique and at a known location $\Delta X^* = c$. This extension is useful to allow for heteroskedasticity in the measurement error. Of course, given the relaxed independence assumption, the assumption of nonvanishing characteristic function of ΔX^* has to be converted to an injectivity assumption regarding the operator $F_{X^*|X}$, as in Hu and Schennach (2008). To handle this extension, one also needs to alter some steps of the proof of Theorem 3.1 that follow Equation (A.11) as follows (we provide a formal proof for the case of one-dimensional X^* — a formal treatment of the multivariate case may be possible at the risk of further complexities):

Alternate proof of Theorem 3.1. (allowing for heteroskedasticity under median restrictions). We now show that $f_{\tilde{X}^*|X}(\tilde{x}^*|x)$ necessarily violates Assumption 2.2 (with ΔX^* replaced by $\tilde{\Delta X}^* \equiv \tilde{X}^* - X$), unless $S(\cdot)$ is the identity function. Let $1(x^* \leq x)$ denote an indicator function of the event $x^* \leq x$ and write:

$$\begin{aligned} \int 1(x^* \leq x + c) f_{X^*|X}(x^*|x) dx^* &= \int 1(x^* \leq x) f_{\Delta X^*|X}(x^* - x|x) dx^* \\ &= \int 1(x^* - x \leq c) f_{\Delta X^*|X}(x^* - x|x) dx^* \\ &= \int 1(u \leq c) f_{\Delta X^*|X}(u|x) du. \end{aligned} \tag{D.1}$$

and by Assumption 2.2 (for a median restriction on ΔX^*), this integral must be equal to $1/2$ for all x for a given known c and for any other value of c this integral differs from $1/2$.

Similarly, if the alternative model is valid we should also have (with \tilde{x}^* such that $x^* = S(\tilde{x}^*)$),

$$\int 1(\tilde{x}^* \leq x + c) f_{\tilde{X}^*|X}(\tilde{x}^*|x) d\tilde{x}^* = 1/2, \tag{D.2}$$

independent of x . Note that under this reparametrization of x^* , the function $h(\cdot)$ in the alternative model (denoted $\tilde{h}(\tilde{x}^*)$), becomes $\tilde{h}(\tilde{x}^*) = h(S(\tilde{x}^*))$. Since $h(x^*)$ is continuous and not constant on any interval by Assumption 3.4 (and therefore so must $\tilde{h}(\tilde{x}^*)$ be if it is to be a valid alternative model), it follows that $S(\tilde{x}^*)$ cannot be discontinuous. Since x^* and \tilde{x}^* are one-dimensional and $S(\cdot)$ is one-to-one (by Assumption 3.3) and continuous, it follows that $S(\cdot)$ must be strictly monotone.

Since $\tilde{x}^* = S^{-1}(x^*)$, the left-hand side of Equation (D.2) can be written as

$$\begin{aligned}
& \int \mathbf{1}(S^{-1}(x^*) \leq x + c) f_{X^*|X}(x^*|x) dx^* \\
&= \int \mathbf{1}(x^* \leq S(x + c)) f_{X^*|X}(x^*|x) dx^* \\
&= \int \mathbf{1}(x^* \leq S(x + c)) f_{\Delta X^*|X}(x^* - x|x) dx^* \\
&= \int \mathbf{1}(x^* - x \leq S(x + c) - x) f_{\Delta X^*|X}(x^* - x|x) dx^* \\
&= \int \mathbf{1}(u \leq S(x + c) - x) f_{\Delta X^*|X}(u|x) du
\end{aligned}$$

where we have assumed that $S(\cdot)$ is strictly monotone *increasing* (a similar treatment holds for $S(\cdot)$ monotone *decreasing*, reversing the direction of the inequality and noting that $1 - 1/2 = 1/2$). By comparison with Equation (D.1), this integral is equal to $1/2$ for any x only if $S(x + c) - x = c$ for any x . This implies that $S(\cdot)$ is the identity function, i.e. the two observationally equivalent models are in fact the same model. \square

Our results cover the case where the measurement equation has a Berkson structure while the instrument equation, though nonlinear, maintains a more classical structure (with the error being independent of the true regressor X^*). It is then natural to wonder whether one could also handle the case where the instrument equation exhibits Berkson-type errors as well, i.e.:

$$\begin{aligned}
X^* &= X + \Delta X^* \\
X^* &= \tilde{h}(Z) + \Delta Z.
\end{aligned}$$

However, it is difficult to find reasonable settings where such a system of equations would be relevant. The two separate “causes” X and Z would have to happen to result in the same “effect” X^* through two different channels with different error terms. In fact, this system of equation implies that the four variables ΔX^* , ΔZ , X , and Z are functionally related through $X + \Delta X^* = \tilde{h}(Z) + \Delta Z$, i.e., one of the four variables is redundant. As a result, this does not constitute a very useful scenario.

E Combination of classical and Berkson errors.

It may also be of interest to consider a combination of a classical and Berkson error resulting in a regression model of the form:

$$\begin{aligned}
Y &= g(X^*) + \Delta Y \\
X^* &= W^* + \Delta X^* \\
X &= W^* + \Delta X
\end{aligned}$$

where Y and X are observed while the other variables are not and where W^* , ΔX^* , ΔX and ΔY are mutually independent. Here, W^* is a Berkson-contaminated measurement of

X^* that is not observed directly. Instead, we observe X , a noisy measure of W^* exhibiting classical errors. Since there are two separate error problems, it is to be expected that we may need two additional pieces of information to identify this model. We could use an observed instrument Z

$$Z = h(X^*) + \Delta Z$$

to address the Berkson aspect of the problem and an observed repeated measurement \tilde{X}

$$\tilde{X} = W^* + \Delta\tilde{X}$$

to address the classical component of the error. More specifically, using the techniques found in Schennach (2004), it is possible to identify the joint distribution of (W^*, Y, Z) from the joint distribution of (X, \tilde{X}, Y, Z) provided $E[\Delta\tilde{X}|W^*, \Delta X] = 0$. The key step is to relate the joint characteristic function $\phi(\gamma, \zeta, \xi)$ of Y, Z, W^* to expectations involving only the observable variables Y, Z, X, \tilde{X} :

$$\begin{aligned} \phi(\gamma, \zeta, \xi) &\equiv E[\exp(\mathbf{i}\gamma Y) \exp(\mathbf{i}\zeta Z) \exp(\mathbf{i}\omega W^*)] \\ &= \frac{E[\exp(\mathbf{i}\gamma Y) \exp(\mathbf{i}\zeta Z) \exp(\mathbf{i}\omega X)]}{E[\exp(\mathbf{i}\omega X)]} \exp\left(\int_0^\omega \frac{\mathbf{i}E[\tilde{X} \exp(\mathbf{i}\xi X)]}{E[\exp(\mathbf{i}\xi X)]} d\xi\right), \end{aligned}$$

where $\mathbf{i} = \sqrt{-1}$. Once the joint distribution of (W^*, Y, Z) is known (by taking the inverse Fourier transform of $\phi(\gamma, \zeta, \xi)$), the techniques introduced in the present paper can be directly used to identify $g(X^*)$, $h(X^*)$ and the densities of X^* , ΔX^* , ΔY , ΔZ . In practice, estimation could be accomplished as before, by writing a suitable likelihood function, integrated over the latent variables, in which the unknown functions are represented by sieve approximations.

References

- BERKSON, J. (1950): “Are there two regressions?,” *Journal of the American Statistical Association*, 45, 164–180.
- BHATTACHARYA, R. N., AND R. R. RAO (2010): *Normal Approximation and Asymptotic Expansions*. SIAM, Philadelphia.
- CARRASCO, M., J.-P. FLORENS, AND E. RENAULT (2005): “Linear Inverse Problems and Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization,” in *Handbook of Econometrics*, vol. Vol. 6. Elsevier Science.
- CARROLL, R. J., X. CHEN, AND Y. HU (2010): “Identification and estimation of nonlinear models using two samples with nonclassical measurement errors,” *Journal of Nonparametric Statistics*, 22, 379–399.
- CARROLL, R. J., A. DELAIGLE, AND P. HALL (2007): “Nonparametric regression estimation from data contaminated by a mixture of Berkson and classical errors,” *Journal of the Royal Statistical Society B*, 69, 859–878.

- CARROLL, R. J., D. RUPPERT, L. A. STEFANSKI, AND C. M. CRAINICEANU (2006): *Measurement Error in Nonlinear Models*. New York: Chapman & Hall.
- DELAIGLE, A., P. HALL, AND P. QIU (2006): “Nonparametric methods for solving the Berkson errors-in-variables problem,” *Journal of the Royal Statistical Society B*, 68, 201–220.
- D’HAULTFOEUILLE, X. (2011): “On the Completeness Condition in Nonparametric Instrumental Problems,” *Econometric Theory*, 27, 460–471.
- DOCKERY, D. W., C. A. POPE, X. P. XU, ET AL. (1993): “An Association Between Air-Pollution And Mortality In 6 United-States Cities,” *New England Journal of Medicine*, 329, 1753–1759.
- DUNFORD, N., AND J. T. SCHWARTZ (1971): *Linear Operators*. John Wiley & Sons, NY.
- FAN, J., AND Y. K. TRUONG (1993): “Nonparametric Regression with Errors in Variables,” *Annals of Statistics*, 21(4), 1900–1925.
- GALLANT, A. R., AND D. W. NYCHKA (1987): “Semi-Nonparametric Maximum Likelihood Estimation,” *Econometrica*, 55, 363–390.
- GINE, E., AND J. ZINN (1990): “Bootstrapping General Empirical Measures,” *The Annals of Probability*, 18, 851–869.
- GRENDER, U. (1981): *Abstract Inference*. Wiley Series, New York.
- 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.
- HU, Y., AND S. M. SCHENNACH (2008): “Instrumental Variable Treatment of Nonclassical Measurement Error Models,” *Econometrica*, 76, 195–216.
- HUWANG, L., AND Y. H. S. HUANG (2000): “On errors-in-variables in polynomial regressions — Berkson case,” *Statistica Sinica*, 10, 923–936.
- HYSLOP, D. R., AND G. W. IMBENS (2001): “Bias from classical and other forms of measurement error,” *Journal of Business & Economic Statistics*, 19, 475–481.
- LEWBEL, A. (1996): “Demand Estimation with Expenditure Measurement Errors on the Left and Right Hand Side,” *The Review of Economics and Statistics*, 78(4), 718–725.
- LI, T. (2002): “Robust and consistent estimation of nonlinear errors-in-variables models,” *Journal of Econometrics*, 110, 1–26.
- LI, T., AND Q. VUONG (1998): “Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators,” *Journal of Multivariate Analysis*, 65, 139–165.

- MAHAJAN, A. (2006): “Identification and Estimation of Single Index Models with Misclassified Regressor,” *Econometrica*, 74, 631–665.
- MALLICK, B., F. O. HOFFMAN, AND R. J. CARROLL (2002): “Semiparametric Regression Modeling with Mixtures of Berkson and Classical Error, with Application to Fallout from the Nevada Test Site,” *Biometrics*, 58, 13–20.
- NELDER, J., AND R. MEAD (1965): “A Simplex Method for Function Minimization,” *Computer Journal*, 7, 308–313.
- NEWBY, W. (2001): “Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models,” *Review of Economics and Statistics*, 83, 616–627.
- NEWBY, W., AND D. MCFADDEN (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by R. F. Engel, and D. L. McFadden, vol. IV. Elsevier Science.
- NEWBY, W. K. (1997): “Convergence rates and asymptotic normality of series estimators,” *Journal of Econometrics*, 79, 147–168.
- NEWBY, W. K., AND J. L. POWELL (2003): “Instrumental Variable Estimation of Nonparametric Models,” *Econometrica*, 71, 1565–1578.
- POPE, C. A., M. J. THUN, M. M. NAMBOODIRI, ET AL. (1995): “Particulate air-pollution as a predictor of mortality in a prospective-study of us adults,” *American Journal of Respiratory and Critical Care Medicine*, 151, 669–674.
- SAMET, J. M., F. DOMINICI, F. C. CURRIERO, ET AL. (2000): “Fine particulate air pollution and mortality in 20 US Cities, 1987-1994,” *New England Journal of Medicine*, 343, 1742–1749.
- SCHENNACH, S. M. (2004): “Estimation of Nonlinear Models with Measurement Error,” *Econometrica*, 72, 33–75.
- (2007): “Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models,” *Econometrica*, 75, 201–239.
- SHEN, X. (1997): “On Methods of Sieves and Penalization,” *Annals of Statistics*, 25, 2555–2591.
- STRAM, D. O., M. HUBERMAN, AND A. H. WU (2002): “Is Residual Confounding a Reasonable Explanation for the Apparent Protective Effects of Beta-carotene Found in Epidemiologic Studies of Lung Cancer in Smokers?,” *American Journal of Epidemiology*, 155, 622–628.
- VAN DER LAAN, M. J., S. DUDOIT, AND S. KELES (2004): “Asymptotic optimality of likelihood-based cross-validation,” *Statistical Applications in Genetics and Molecular Biology*, 3, 4.

WANG, L. (2004): “Estimation of nonlinear models with Berkson measurement errors.,”
Annals of Statistics, 32, 2559–2579.

——— (2007): “A unified approach to estimation of nonlinear mixed effects and Berkson
measurement error models,” *Canadian Journal of Statistics*, 35, 233–248.