Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: www.tandfonline.com/journals/gnst20

Identification and estimation of nonlinear models using two samples with nonclassical measurement errors

Raymond J. Carroll, Xiaohong Chen & Yingyao Hu

To cite this article: Raymond J. Carroll, Xiaohong Chen & Yingyao Hu (2010) Identification and estimation of nonlinear models using two samples with nonclassical measurement errors, Journal of Nonparametric Statistics, 22:4, 419-423, DOI: 10.1080/10485250903556110

To link to this article: https://doi.org/10.1080/10485250903556110

Published online: 30 Apr 2010.

Journal of Nonparametric Statistics
Vol. 22, No. 4, May 2010, 419–423

Taylor & Francis
Taylor & Francis Group

REJOINDER

Identification and estimation of nonlinear models using two samples with nonclassical measurement errors

Raymond J. Carroll[a]*, Xiaohong Chen[b] and Yingyao Hu[c]

[a]*Department of Statistics, Texas A & M University, 447 Blocker Building, College Station, TX 77843-3143, USA;* [b]*Department of Economics, Yale University, Box 208281, New Haven, USA;* [c]*Department of Economics, Johns Hopkins University, 440 Mergenthaler Hall, 3400 N. Charles Street, Baltimore, MD 21218, USA*

(*Received 5 December 2009; final version received 11 December 2009*)

We are grateful to the discussants for their insightful comments. Below is our rejoinder and clarification related to their main discussions.

The comments from Aurore Delaigle and Peter Hall mainly focus on two issues: the availability of the auxiliary sample satisfying the assumptions and the empirical choice of smoothing parameters. Our key identification assumptions require that the conditional distribution, $f_{Y|X^*,W}$, of the dependent variable Y given (X^*, W) is the same in both samples, but that the distribution, $f_{X^*|W}$, of the independent variables differs in the two samples. We think these are reasonable in many applications. For example, we consider a consumption model, where Y is an individual's consumption, X^* an individual's true log income and W the gender and marital status. We use a primary sample from the current population survey (CPS) and an auxiliary sample from the survey of income and program participation (SIPP). Generally speaking, the CPS is from the whole US population, while the SIPP focuses more on participants in welfare programmes with more welfare-related questions. Most economic theory models assume that the conditional distribution $f_{Y|X^*,W}$ is the same in both samples. It is clearly reasonable to believe that the latent log income distribution $f_{X^*|W}$ for given gender and marital status is different from the support, that is, the whole real line, in the two samples. An advantage of our results is that we allow the misreporting errors to have different distributions in the two samples.

The empirical choice of smoothing parameters is always a difficult task in semi/nonparametric estimation. This is partly why, after establishing nonparametric identification of $f_{Y|X^*,W}$, $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$, and $f_{X_a^*|W_a}$, we consider semiparametric estimation by specifying $f_{Y|X^*,W} = g(Y|X^*, W; \theta_0)$ parametrically up to a finite-dimensional unknown parameter of interest θ_0, and treating $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a}$ as nuisance functions. In a closely related problem of sieve minimum distance (MD) estimation of semi/nonparametric ill-posed inverse instrumental

*Corresponding author. Email: carroll@stat.tamu.edu

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250903556110
http://www.informaworld.com

variables regression models, Chen and Pouzo (2009) have shown that the same choice of sieve number of terms can simultaneously lead to the optimal convergence rate of a sieve MD estimator of the nonparametric nuisance functions and the root-n asymptotic normality of a sieve MD estimator of the finite-dimensional parameter of interest. This nice property only holds for semi/nonparametric models estimated via the method of sieves, and is not valid for kernel-based semi/nonparametric estimators in general. It is very easy to show that this nice property also holds for the sieve maximum-likelihood (ML) estimation of semi/nonparametric measurement error models. Consequently, we could use various existing data-driven methods and/or model selection criteria such as modified AIC and BIC to select the sieve number of terms for the nonparametric nuisance functions, and the resulting estimator of θ will still be root-n asymptotically normal.

Delaigle and Hall also comment on the choice of sieve bases in the section on semiparametric estimation. They mention that the particular sieve basis approximation we suggested could be viewed as a slightly more general version of the classical measurement error model. It is true that one should be more careful with the choice of sieve approximation bases when the main parameter of interest is the latent probability conditional and/or marginal densities. However, if one cares mainly about estimation and inference of the finite-dimensional parameter θ associated with the parametrically specified conditional density $f_{Y|X^*,W}$, then our experience indicates that the choice of sieve bases is not important. Moreover, as the sieve number of terms grows to infinity with sample size n, even this simple sieve basis suggested in Section 3 can still approximate any unknown density functions. Of course many other sieve basis approximation could also be used.

Marie-Luce Taupin provides a nice brief review of the recent statistical literature on density deconvolution and nonlinear regression models for classical measurement error problems. She also presents concerns about the identification and estimation results in our paper. As an example, she considers a simple case of identification, where

$$Y = \theta_0 X^* + \xi, \quad X = X^* + \varepsilon,$$

and X^*, ξ and ε are mutually independent. Her concern is that the model is not identified when $\theta_0 = 0$. In other words, the operator $L_{X,Y}$ cannot be injective in this case. This setting violates our Assumption A.1, hence our Theorem 2.1 is not applicable. Our Assumption A.1 does require the latent explanatory variable X^* to be relevant in the model. In fact, it requires the dependent variable Y to be informative enough about the latent variable X^* in the sense that the operator $L_{X,Y}$ is injective. For example, the injectivity assumption would also fail when we have

$$Y = I(\theta_0 X^* + \xi > 0), \quad X = X^* + \varepsilon,$$

where $I(\cdot)$ is the indicator function. In this example, even if $\theta_0 \neq 0$, a discrete-dependent variable Y is not informative enough for a continuous latent variable X^* since the operator $L_{X,Y}$ is not injective. This example is also ruled out by our Assumption A.1.

Although our paper provides nonparametric identification of $f_{Y|X^*,W}$, $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a}$, we consider a semiparametric estimation by specifying $f_{Y|X^*,W}$ parametrically, $g(Y|X^*, W; \theta_0)$, up to a finite-dimensional unknown parameter θ_0, and treating $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a}$ as nuisance functions. It is true that to obtain the optimal convergence rates in weighted sup-norm or $\mathcal{L}^2$-norm for sieve estimators of $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a}$ is a difficult ill-posed inverse problem. In particular, the problem could be *severely* ill-posed in the sense that some of these density functions are estimated at a $\log(n)$ rate in a weighted sup-norm or $\mathcal{L}^2$-norm. For example, if the measurement error regression model in the primary sample is a classical one, that is, $f_{X|X^*} = f_\varepsilon(x - x^*)$ with a Gaussian measurement error ε, then for ordinarily smooth density function $f_{X^*|W}$, for example, $f_{X^*|W}$ belongs to a Holder, Sobolev or Besov space with finite smoothness, the convergence rate of its sieve estimator in weighted sup-norm or $\mathcal{L}^2$-norm will be of a logarithmic order. Hence, the finite sample performance of sieve estimators of $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a}$ is typically more sensitive to the choices of sieve bases and sieve number of terms than that of sieve estimators for direct density estimation problems, that is, when X^* are observable. This is partly why we focus on estimation of the finite-dimensional parameter θ, and treating $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a}$ as nuisance functions. Our experience is that to obtain root-n asymptotic normality estimation of θ, the choices of sieve bases and sieve number of terms are not very crucial. As clearly shown in Van der Vaart (1996), Ai and Chen (2003) and Chen and Pouzo (2009), for root-n asymptotic normality estimation of θ in semi/nonparametric ill-posed inverse problems, what matters is to choose the sieves and sieve number of terms to ensure faster than $n^{-1/4}$ rates of convergence of sieve estimators of these nuisance functions under the Fisher metric, even though the rate in $\mathcal{L}^2$-norm is slower than $n^{-1/4}$. Applying the general theory on semiparametric sieve ML estimation of Shen (1997) and on sieve MD estimation of semi/nonparametric ill-posed inverse problems of Chen and Pouzo (2009), we know that the nonparametric convergence rates of our sieve estimators of $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a}$ depend on the sieve approximation bias order as well as complexity of our semiparametric mixture models. For our semiparametric measurement error regression model, under the assumption that $f_{X|X^*}$, $f_{X^*|W}$, $f_{X_a|X_a^*}$ and $f_{X_a^*|W_a}$ belong to a Holder, Sobolev or Besov space with finite smoothness, the optimal convergence rates of sieve estimators of these density functions in weighted sup-norm or $\mathcal{L}^2$-norm are typically slower than $n^{-1/4}$, but the corresponding convergence rates in Fisher norm is faster than $n^{-1/4}$. For *linear* ill-posed inverse problems, even when the nonparametric estimation in $\mathcal{L}^2$-norm is severely ill-posed, a nonparametric rate faster than $n^{-1/4}$ in Fisher norm is sufficient to ensure root-n asymptotic normality of estimators of θ. Unfortunately, for *nonlinear* ill-posed inverse problems, our current sufficient conditions to ensure root-n asymptotic normality of estimators of θ rules out severely ill-posed cases. See Chen and Pouzo (2009) for explanation as to why one may fail to obtain root-n estimation of θ for *nonlinear severely* ill-posed problems.

Young Truong is mainly concerned with the existence of an optimum solution of the sieve parameters β in sieve ML estimation. There are two kinds of parameters in our sieve ML estimation of the semiparametric measurement error model specified in our Section 3: the parameter of interest θ used to specify the conditional density $f_{Y|X^*,W}$ parametrically as $g(y|x^*, w; \theta)$ and the sieve parameters β associated with the sieve bases used to approximate the unknown nuisance functions such as $f_{X|X^*}$ and $f_{X^*|W}$. Our Assumption 3.2(iii) requires that the parametrically specified latent conditional density model $g(y|x^*, w; \theta)$ is identified when there were no measurement errors involved. That is, our Assumption 3.2(iii) requires $\int[\log g(y|x^*, w; \theta)] f_{Y|X^*,W}(y|x^*, w)\mathrm{d}y$ to have a unique maximiser $\theta_0 \in \Theta$, where Θ is a compact subset of $\mathbb{R}^{d_\theta}$ and d_θ is finite and independent of sample size n. This assumption allows us to focus on the identification problem caused by measurement errors. It is true that without any conditions the local optimum solution of θ may not exist in the presence of measurement errors. Our paper provides a set of reasonable conditions so that all the latent density functions $f_{Y|X^*,W}$, $f_{X|X^*}$ and $f_{X^*|W}$ are nonparametrically identified. These nonparametric identification results and our Assumption 3.2(iii) imply that the expectation of the log-likelihood function of the data will be maximised at the true θ_0, $f_{X|X^*}$ and $f_{X^*|W}$. Therefore, the sieve ML does have a unique maximiser $(\hat{\theta}, \hat{f}_{X|X^*}, \hat{f}_{X^*|W})$ when sample size is reasonably large. Nevertheless, once the unknown nuisance functions $f_{X|X^*}$ and $f_{X^*|W}$ are approximated by finite-dimensional sieves with sieve parameters β, it is possible that a sieve ML does not have a unique maximiser in terms of sieve parameters β for a particular sieve basis. This is not too surprising since different combinations of sieve bases could approximate the same true unknown solutions $f_{X|X^*}$ and $f_{X^*|W}$ and hence it is likely that although the solutions $\hat{f}_{X|X^*}$ and $\hat{f}_{X^*|W}$ are well defined their corresponding sieve parameter estimators $\hat{\beta}$ are not unique.

The asymptotic properties of the sieve ML estimator $(\hat{\theta}, \hat{f}_{X|X^*}, \hat{f}_{X^*|W})$ only assumes that the estimator is well defined and does not require uniqueness of the individual sieve parameter estimator $\hat{\beta}$ in any given sample. Of course, if one also wants to have a unique optimum $\hat{\beta}$ in any given sample, then one should be more careful in terms of the choice of sieve bases. Finally, our nonparametric identification strategy is constructive and one could propose an alternative estimation procedure by following the identification proofs.

Han Hong provides a nice brief review of the recent econometrics literature on nonclassical measurement error models prior to 2006, and also presents a fair comparison between our paper and Chen, Hong, and Tamer (2005). As for the identification literature, there are also studies an statistics which provide constructive identification of nonlinear models with a misclassified binary variable. For example, Hui and Walter (1980) show that models with a latent binary regressor can be nonparametrically expressed as explicit functions of observed probabilities when two conditionally independent measurements of the latent variable are available.

Hong nicely suggests that one could compare the semiparametric efficiency gain of the sieve MLE estimator $\hat{\theta}$ when both the assumptions of our paper and those of Chen et al. (2005) hold, which could then lead to a test of the validity of the existence of a validation sample, that is, the secondary sample contains observation of the true value X_a^*. Since we may identify the error distributions in both samples, it is feasible to test whether there are measurement errors in the secondary sample. When the secondary sample is a validation sample, the estimator in Chen et al. (2005) should be more efficient than the one in our paper because their estimator uses the information that the true values X_a^* are observed in the secondary sample, and their estimator of θ is always root-n asymptotically normal regardless of what kind of measurement errors occur in the primary sample. Although we provide sufficient conditions to ensure root-n asymptotic normality of our sieve MLE of θ when both samples contain possibly nonclassical measurement errors and the true value X_a^* is never observed, our current sufficient conditions rule out *nonlinear* measurement error problems that are *severely* ill-posed.

Hong also suggests to test some of our conditions imposed for nonparametric identification. It is true that some of our assumptions, such as Assumption 2.3, are testable from the data under other identification assumptions. A special case which violates Assumption 2.5 is that $f_{X^*|W_j} = f_{X^*|W_i}$ and $f_{X_a^*|W_j} = f_{X_a^*|W_i}$, which holds if and only if the observed densities satisfy $f_{X|W_j} = f_{X|W_i}$ and $f_{X_a|W_j} = f_{X_a|W_i}$ under Assumptions 2.1, 2.2 and 2.4. Therefore, this special case is directly testable from the data. However, it would be very difficult to test all the identification assumptions at the same time. The invertibility Assumption A.1 is directly testable when the latent variable X^* is discrete and the linear operators are in fact matrices. For a continuous X^*, testing the invertibility of an operator means to test whether the corresponding nonparametric distribution is complete, which is quite challenging.

Our identification strategy assumes that the perfectly measured covariates W are discrete. When they are continuous, the identification results still hold with some modification of the assumptions. To be specific, Assumption 2.1(ii) implies that

$$f_{X,Y,W}(x, y, w) = \int f_{X|X^*}(x|x^*) f_{Y,X^*,W}(y, x^*, w)\mathrm{d}x^*,$$

where W may be continuous. Discretising W to W^d leads to

$$f_{X,Y|W^d}(x, y|w_j^d) = \int f_{X|X^*}(x|x^*) f_{Y|X^*,W^d}(y|x^*, w_j^d) f_{X^*|W_j^d}(x^*)\mathrm{d}x^*,$$

which is the same as Equation (A2) in the current setting. Although the density $f_{Y|X^*,W^d}$ is no longer the model of interest $f_{Y|X^*,W}$, the same identification strategy may still pin down the error distribution $f_{X|X^*}$. Since the operator corresponding to $f_{X|X^*}$ is injective, the density $f_{Y,X^*,W}$

may still be identified from the observed $f_{X,Y,W}$. Therefore, the identification with continuous covariates is feasible under certain modification of the current assumptions.

Acknowledgements

Chen acknowledges support from the National Science Foundation (SES-0631613). Carroll's research was supported by grants from the National Cancer Institute (CA57030, CA104620), and partially supported by Award Number KUS-CI-016-04 made by King Abdullah University of Science and Technology (KAUST).

References

Ai, C., and Chen, X. (2003), 'Efficient Estimation of Conditional Moment Restrictions Models Containing Unknown Functions', *Econometrica*, 71, 1795–1843.
Chen, X., and Pouzo, D. (2009), 'Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals', *Journal of Econometrics*, 152, 46–60.
Chen, X., Hong, H., and Tamer, E. (2005), 'Measurement Error Models with Auxiliary Data', *Review of Economic Studies*, 72, 343–366.
Hui, S.L., and Walter, S.D. (1980), 'Estimating the Error Rates of Diagnostic Tests', *Biometrics*, 36, 167–171.
Shen, X. (1997), 'On Methods of Sieves and Penalization', *Annals of Statistics*, 25, 2555–2591.
Van der Vaart, A. (1996), 'Efficient Maximum Likelihood Estimation in Semiparametric Mixture Models', *Annals of Statistics*, 24, 862–878.