

NONPARAMETRIC IDENTIFICATION OF LATENT COMPETING RISKS MODELS

GORDANA COLBY
York University

PAUL RILSTONE
York University

This paper shows that nonparametric identification of latent competing risks models is possible without the usual conditional independence and exclusion restrictions.

1. INTRODUCTION

The identification of competing risks models has been addressed by a number of authors. Cox (1962) and Tsiatis (1978) show that, in the absence of independence, the underlying joint distribution is, in general, not identifiable. Heckman and Honoré (1989, 1990) and Abbring and van den Berg (2003) demonstrate that, with observable covariates, identification of these models is possible. Omori (1998) shows identifiability of independent competing risks with multiple spells. Heckman and Honoré (1990) allow for latent situations where one observes the maximum of two processes but does not know which one was observed. Examples of these from epidemiology are in Lee (1992), from reliability analysis in Meeker and Escobar (1998), and from health economics in Paric and Rilstone (2000).

The processes in Heckman and Honoré (1989, 1990) are dependent in that both are functions of unobservables. Identification is obtained using covariate exclusion restrictions. Conditional on the covariates and the unobservables the processes are independent. This excludes situations where one wants to consider the direct interaction between durations. Furthermore, in certain circumstances it may not be plausible to employ exclusion restrictions. This paper extends the identification results in Heckman and Honoré (1989, 1990) by allowing for conditional dependence between the two latent processes and allowing for the processes to be conditional upon the same set of covariates.

The authors thank Jinyong Hahn, James Heckman, and Shinichi Sakata for useful discussions on the subject matter of this paper. Any errors are the fault of the authors. Research funding for Rilstone was provided by the Social Sciences and Humanities Research Council of Canada. Address correspondence to Gordana Colby or Paul Rilstone, Department of Economics, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3.

One way of seeing the identification problem here is that the observed dependent variables are sample order statistics (the minimum in this case) from repeated samples whose individual outcomes are unobserved. Clearly, the results we derive can be applied to situations where only the maximum of repeated samples is observed or, in fact, any (known) order statistics. As we discuss in Section 4, the results we develop can be adapted to a number of cases such as sample selection and disequilibrium models.

As is standard in this literature, the results are nonparametric and allow for unobservable frailty or heterogeneity. For clarity, we assume throughout that there are only two processes, although this can be generalized.

The discussion proceeds as follows. In Section 2 we allow for direct dependence among the risk sets and show when identification is possible. In Section 3 we show how identification can be obtained even without exclusion restrictions. In each of these cases, the results provide sufficient conditions for identification.

2. IDENTIFICATION WITH CONDITIONALLY DEPENDENT DURATION VARIABLES

The Heckman and Honoré (1989) competing risks framework can be generalized as follows. Let U_1 and U_2 be unobservables on $(0,1) \times (0,1)$ with joint distribution and density functions K and k and let $x = (x_1, x_2, x_3, z)$ be a vector of covariates taking values on $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \times \mathcal{Z}$. Recursively define two duration processes, T_1 and T_2 :

$$\Phi_1(T_1) = -\frac{\log U_1}{\theta_1(x_1, z)}, \quad \Phi_2(T_2) = -\frac{\log U_2}{\theta_2(x_2, z)\rho(T_1, x_3, z)}, \tag{1}$$

where $\Phi_j, j = 1, 2$ are commonly referred to as the cumulative baseline hazard rates. As discussed in Heckman and Honoré (1989), special cases of this include the proportional and accelerated hazard rate models. This framework is also applicable to multiple spell models and various nonseparable hazard rate models. We index the x 's to allow for exclusion restrictions such that x_j appears only in $\theta_j, j = 1, 2$. The term x_3 appears only in ρ , whereas z can be common to both hazard rates.¹ This is similar to exclusion restrictions in simultaneous equations estimation. Put $\psi = (\theta_1, \theta_2, \rho)$. Let

$$\Lambda_1 \equiv \Phi_1(t_1)\theta_1(x_1, z), \quad \Lambda_{2|1} \equiv \Phi_2(t_2)\theta_2(x_2, z)\rho(t_1, x_3, z). \tag{2}$$

By the transformation theorem, the density and survivor functions of T_1, T_2 are derived as

$$f(t_1, t_2|x) = |J|k(e^{-\Lambda_1}, e^{-\Lambda_{2|1}}) \quad \text{and} \quad S(t_1, t_2|x) = K(e^{-\Lambda_1}, e^{-\Lambda_{2|1}}), \tag{3}$$

where

$$|J| = \Phi'_1 \theta_1 \Phi'_2 \theta_2 \rho e^{-\Lambda_1} e^{-\Lambda_2} \tag{4}$$

and Φ'_j denotes the derivative of Φ_j . We let $\mathfrak{R}_+ = [0, \infty)$.

We formalize our assumptions as follows.

Assumption 1. The durations are generated as in equation (1) where $(U_1, U_2) \sim K : (0,1) \times (0,1) \rightarrow [0,1]$, $K(\cdot, \cdot)$ is strictly increasing and differentiable in its arguments.

Assumption 2. For $j = 1, 2$, $\Phi_j(t)$ is differentiable and strictly increasing, $\Phi_j(0) = 0$, $\Phi_j(1) = 1$, and $\Phi_j(\infty) = \infty$.

Assumption 3. $(\theta_1(x_1, z), \theta_2(x_2, z)) \rightarrow R_+^2$ is continuous. For $j = 1, 2$, some $x_j^1 \in \mathcal{X}_j$ and $x_j^0 \in \mathcal{X}_j$, $\theta_j(x_j^1, z) = 1$, and $\theta_j(x_j^0, z) = \lim_{x_j \rightarrow x_j^0} \theta_j(x_j, z) = 0$.

Assumption 4. $\rho(t, x_3, z) \rightarrow (a, b)$, $a < 1 < b$, is differentiable with respect to t , $\rho(0, x_3, z) = 1$, and for some $x_3^1 \in \mathcal{X}_3$, $\rho(t, x_3^1, z) = 1$; $\psi^{-1}(\mathfrak{R}_+^2 \times (a, b)) = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \times \mathcal{Z} \times \mathfrak{R}_+$.

The statements regarding the domains, ranges, and inverse images of θ and ρ are made to ensure that, when evaluating one of the variates at a fixed point, this does not inadvertently impose restrictions on the values the other variates may take. The intuition of the results is very simple. In a competing risks framework, the observed duration variable is the minimum of the two duration processes. We find conditions under which the ‘‘observable’’ survivor function tends to the marginal survivor function of the processes. This kind of approach is sometimes referred to as ‘‘identification at infinity’’ as in Heckman (1990). Essentially, the assumption is made that, for some value of the conditioning variables (not necessarily infinity) or set of values, one of the dependent variables has a degenerate distribution. Given that, results like those of Elbers and Ridder (1982) can be applied directly. Moreover, because the parameters of interest are expressed directly in terms of the survivor function and its derivatives, these can be estimated directly by their sample analogs.

The partitioning of $x = (x_1, x_2, x_3, z)$ allows for a subset of the observed covariates to enter into ρ . We show identification of this model under two different sets of assumptions. We either assume that x_3 is a subset of x with no elements in common with either x_1, x_2 , or z and x_3 has a certain limiting impact on ρ , or we will show that certain shape restrictions on Φ_2 and ρ can identify the components of this model. Under the assumption that $\rho(t, x_3^1, z) = 1$, T_1 has no direct impact on T_2 and the duration processes become conditionally independent. In this situation identification essentially follows along the lines of Heckman and Honoré (1989, 1990). We impose the natural restriction that $\rho(0, x_3, z) = 1$ so that the marginal distributions with this model are also as in the conditionally independent model.

We let $K_1(u_1) = K(u_1, 1)$, $K_2(u_2) = K(1, u_2)$, denote the marginal distribution functions and indicate their derivatives by K'_j , $j = 1, 2$. The partial derivatives of K are indicated by $K^{(j)}$, $j = 1, 2$. Note that $K_j(0) = 0$ and $K_j(1) = 1$, $j = 1, 2$.

Assumption 5. For $j = 1, 2$ K'_j is left continuous at 1 and $K'_j(1) = L_j > 0$.

Assumption 5 guarantees that a certain limiting ratio of the K'_j 's in the proofs that follow is well defined. Alternatively, one could allow for the K'_j 's to vanish, by, say, strengthening the smoothness assumption and using L'Hôpital's rule.

PROPOSITION 1. *Let Assumptions 1–5 hold. Then, θ_j , Φ_j , $j = 1, 2$, ρ , and K are identifiable.*

Proof. The survivor function of the minimum of the two durations is given by $S(t|x) = K(e^{-\Lambda_1(t|x)}, e^{-\Lambda_2(t|x)})$. (5)

We can identify θ_2 by evaluating the ratio

$$\frac{\frac{\partial}{\partial t} S(t|x_1^0, x_2, x_3, z)}{\frac{\partial}{\partial t} S(t|x_1^0, x_2^1, x_3, z)} = \frac{\theta_2(x_2, z) K'_2(e^{-\Phi_2(t)\theta_2(x_2, z)\rho(t, x_3, z)}) \frac{\partial}{\partial t} (\Phi_2(t)\rho(t, x_3, z))}{K'_2(e^{-\Phi_2(t)\rho(t, x_3, z)}) \frac{\partial}{\partial t} (\Phi_2(t)\rho(t, x_3, z))} = \theta_2(x_2, z) \frac{K'_2(e^{-\Phi_2(t)\theta_2(x_2, z)\rho(t, x_3, z)})}{K'_2(e^{-\Phi_2(t)\rho(t, x_3, z)})}$$
 (6)

and observing that

$$\theta_2(x_2, z) = \lim_{t \rightarrow 0} \frac{\frac{\partial}{\partial t} S(t|x_1^0, x_2, x_3, z)}{\frac{\partial}{\partial t} S(t|x_1^0, x_2^1, x_3, z)}$$
 (7)

We can identify θ_1 symmetrically. To identify K note that

$$S(1|x_1, x_2, x_3^1, z) = K(e^{-\theta_1(x_1, z)}, e^{-\theta_2(x_2, z)})$$
 (8)

By varying θ_1 and θ_2 over \mathfrak{N}_+^2 , we can trace out K . Because K is identifiable and increasing in its arguments, K_2 has a unique and identifiable inverse, H_2 , such that

$$e^{-\Phi_2(t)\theta_2(x_2, z)\rho(t, x_3, z)} = H_2(S(t|x_1^0, x_2, x_3, z)),$$
 (9)

so that

$$\Phi_2(t) = -\ln(H_2(S(t|x_1^0, x_2^1, x_3^1, z)))$$
 (10)

and

$$\rho(t, x_3, z) = \frac{\ln(H_2(S(t|x_1^0, x_2^1, x_3, z)))}{\ln(H_2(S(t|x_1^0, x_2^1, x_3^1, z)))}. \tag{11}$$

The term Φ_1 is identified symmetrically. ■

Remark 1. Without the identifying variable, x_3 , it is still possible to identify the parameters using functional form restrictions. Note that in this case we can identify

$$\rho(t)\Phi_2(t) = -\ln(H_2(S(t|x_1^0, x_2^1))). \tag{12}$$

In general, it is not possible to decompose the left-hand side into the two functions. However, for certain, fairly rich parametric models one can show that, if $\rho(t) = \rho(t; a)$ and $\Phi(t) = \Phi(t; b)$, where the true values are a_0 and b_0 , say, then for any other values, a' and b' , say, $\rho(t; a_0) \neq \rho(t; a')$ and $\Phi(t; b_0) \neq \Phi(t; b')$ on a set of positive measure. Consequently, it is possible to identify the two functions.

For example, suppose $\Phi(t) = t^a$, $\rho(t) = (1 + t)^b$ and say for two values a_0, b_0 and a', b' these are the same. Then

$$\begin{aligned} 0 &= \frac{d}{dt} (a_0 \log t + b_0 \log(1 + t)) - \frac{d}{dt} (a' \log t + b' \log(1 + t)) \\ &= \frac{a_0 - a'}{t} + \frac{b_0 - b'}{1 + t} \end{aligned} \tag{13}$$

or

$$t = -\left(\frac{1}{a_0 - a'} + \frac{1}{b_0 - b'}\right)(b_0 - b'), \tag{14}$$

a constant, which can only hold on a set of measure zero. The rest of the identification follows as previously.

3. IDENTIFICATION WITHOUT EXCLUSION RESTRICTIONS

In many situations it may not be plausible to impose exclusion restrictions. For example, in bargaining situations, if both agents have access to the same information, there is no reason to expect that one agent will condition on less information than the other. However, it may be plausible to make an assumption as to how a covariate will impact on each agent’s duration dependence, at least in a limiting sense. In this case the data are generated as

$$\Phi_1(T_1) = -\frac{\log U_1}{\theta_1(x, z)}, \quad \Phi_2(T_2) = -\frac{\log U_2}{\theta_2(x, z)}, \tag{15}$$

where x (a scalar) and z , taking values on $\mathcal{X} \times \mathcal{Z}$, are common to both θ_1 and θ_2 . The durations respond asymmetrically to x such that for, say, large (small) values of x , we have $\theta_1 \rightarrow 0$ ($\theta_2 \rightarrow 0$) and we can assume that an observed exit is due to T_2 (T_1) being the minimum. It should be intuitive how the model is identified. The terms Φ_2 and θ_2 are identified using values of x such that, say, $x \geq x^{10}$. Conversely, Φ_1 and θ_1 are identified using values of x such that, say, $x \leq x^{20}$ ($x^{20} < x^{10}$). We require that the relevant normalizations can be imposed within these ranges. The assumptions are as follows.

Assumption 6. The durations are generated as in equation (15) where $(U_1, U_2) \sim K: (0, 1) \times (0, 1) \rightarrow [0, 1]$, $K(\cdot, \cdot)$ is strictly increasing and differentiable in its arguments.

Assumption 7. $\theta(x, z) = (\theta_1(x, z), \theta_2(x, z)) \rightarrow R_+^2$ is continuous; $\theta^{-1}(\mathfrak{R}_+^2) = \mathcal{X} \times \mathcal{Z}$.

Assumption 8. For all $(x, z) \in \mathcal{X} \times \mathcal{Z}$ such that $x > x^{10}$, $\theta_1(x, z) = 0$ and for all $(x, z) \in \mathcal{X} \times \mathcal{Z}$ such that $x < x^{20}$, $\theta_2(x, z) = 0$, some $x^{10} > x^{20}$. For some $(x^2, z^2) \in \mathcal{X} \times \mathcal{Z}$, $x^2 > x^{10}$, $\theta_2(x^2, z^2) = 1$; for some $(x^1, z^1) \in \mathcal{X} \times \mathcal{Z}$, $x^1 < x^{20}$, $\theta_1(x^1, z^1) = 1$.

PROPOSITION 2. *Let Assumptions 2 and 5–8 hold. Then Φ_j , $j = 1, 2$ and K are identifiable, θ_1 is identifiable for $x < x^{20}$, and θ_2 is identifiable for $x > x^{10}$.*

Proof. Note first that for values of x such that $x > x^{10}$

$$\frac{\frac{\partial}{\partial t} S(t|x, z)}{\frac{\partial}{\partial t} S(t|x^2, z^2)} = \frac{\Phi_2'(t)\theta_2(x, z)K_2'(\Phi_2(t)\theta_2(x, z))}{\Phi_2'(t)K_2'(\Phi_2(t))} = \theta_2(x, z) \frac{K_2'(\Phi_2(t)\theta_2(x, z))}{K_2'(\Phi_2(t))}, \tag{16}$$

so that

$$\theta_2(x, z) = \lim_{t \rightarrow 0} \frac{\frac{\partial}{\partial t} S(t|x, z)}{\frac{\partial}{\partial t} S(t|x^2, z^2)}. \tag{17}$$

We identify $\theta_1(x, z)$ symmetrically. Setting $t = 1$ we can identify K by varying θ_1 and θ_2 over \mathfrak{R}_+^2 . As in the proof of Proposition 1, we can then identify Φ_j by inverting K_j , $j = 1, 2$. ■

Remark 2. Note that Proposition 2 only partially identifies the model (at least nonparametrically). Without imposing other restrictions we can identify $\theta_1(x, z)(\theta_2(x, z))$ only for those values of x, z such that $x < x^{20}$ ($x > x^{10}$).

Remark 3. The presence of z in θ_1 and θ_2 allows us to vary these functions over \mathfrak{R}_+^2 and identify K . Without z we would only be able to observe θ_1 and θ_2 over, say, $(x^{20}, \infty) \times (x^{10}, \infty)$, so that K and $\Phi_j, j = 1, 2$, are only partially identifiable. This problem can be partially circumvented by assuming θ_1, θ_2 are differentiable. An explanation is in the working paper version of this paper (Paric and Rilstone, 1999).

4. CONCLUSION

This paper has shown that, under a variety of restrictions, a class of latent competing risks models can be identified. Because the functions we consider are identified in terms of the observed durations and their derivatives, our results suggest a couple of estimation strategies. One possibility would be to use nonparametric analogs of the survivor function and its derivatives. There are both theoretical and practical difficulties involved with this, because the estimation would involve calculations at boundary points. Alternatively, one could parameterize the components of the model and use standard maximum likelihood procedures. Simulation and empirical results in Paric and Rilstone (1998, 2000) indicate that this works quite well.

The identification results in the paper can be readily extended to a number of other econometric models. In the case of sample selection problems one often observes the minimum or maximum of two or more random variables. The prototype of this is the Roy (1951) model where one observes the maximum of, say, salary offers. The framework of this paper is readily adaptable to that model. Moreover, it is quite intuitive that the results would be useful in the Roy context if an agent conditions on a sequence of wage offers. The same analysis applies to auction models.

The latent feature of the data that is central to our paper is also present in disequilibrium models of supply and demand where one observes the minimum of these two functions but does not know which has been observed. A number of examples are given in Maddala (1986). Our results are directly applicable with a simple reinterpretation of the variables and functions in the model.

Another extension of the results of this paper would be for an intermediate situation along the lines of Lee and Porter (1984) in which the “cause of failure” is imperfectly observed. In this case, additional observables would be DY and Y , say, where D indicates which process was the minimum and Y equals one if D is observed, zero otherwise, and possibly depends on the covariates. Because Y is observable, its distribution is identifiable. The complete observations could be used to identify the distribution of the rest of the process.

NOTE

1. It is tempting to generate T_1 and T_2 symmetrically so that a function, say, $\rho_1(T_2)$, appears in the denominator of the first term in equation (1). In fact, one could do this and mechanically walk through the identification proofs that follow. However, because of the inherent nonlinearities involved in equation (1), T_1 and T_2 would not be uniquely defined except under rather implausible conditions.

REFERENCES

- Abbring, J.H. & G.J. van den Berg (2003) The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society Series B* 65, 701–710.
- Cox, D.R. (1962) *Renewal Theory*. Methuen.
- Elbers, C. & G. Ridder (1982) True and spurious duration dependence: The identifiability of the proportional hazard model. *Review of Economic Studies* 49, 403–410.
- Heckman, J.J. (1990) Varieties of selection bias. *American Economic Review*, Papers and Proceedings, 313–318.
- Heckman, J.J. & B.E. Honore (1989) The identifiability of the competing risks model. *Biometrika* 76, 325–330.
- Heckman, J.J. & B.E. Honore (1990) The empirical content of the Roy model. *Econometrica* 58, 1121–1149.
- Lee, E.T. (1992) *Statistical Methods for Survival Data Analysis*. Wiley.
- Lee, L.F. & R.H. Porter (1984) Switching regression models with imperfect sample separation information: With an application on Cartel stability. *Econometrica* 52, 391–418.
- Maddala, G.S. (1986) Disequilibrium, self-selection, and switching models. In Z. Griliches & M.D. Intriligator (eds.), *Handbook of Econometrics*, vol. 3, pp. 1633–1688. North-Holland.
- Meeker, W.Q. & L.A. Escobar (1998) *Statistical Methods for Reliability Data*. Wiley.
- Omori, Y. (1998) The identifiability of independent competing risks models with multiple spells. *Oxford Bulletin of Economics and Statistics* 60, 107–116.
- Paric, Gordana & Paul Rilstone (1998) Structural Estimation of Psychiatric Hospital Stays. Working paper, York University.
- Paric, Gordana & Paul Rilstone (1999) Nonparametric Identification of Latent Competing Risks and Roy Duration Models. Working paper, York University.
- Paric, Gordana & Paul Rilstone (2000) Simplified Estimation of Multivariate Duration Models with Unobserved Heterogeneity. Working paper, York University.
- Roy, A.D. (1951) Some thoughts on the distribution of earnings. *Oxford Economic Papers*, n.s., 3, 135–146.
- Tsiatis, A. (1978) A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences* 72, 20–22.