Nonparametric Identification*

Katherine Hauck and Tiemen Woutersen

March 12, 2024

Keywords: Nonparametric Identification, Parametric Identification, Analysis of Variations

1 Introduction

Economic models can have structural unobserved heterogeneity, as well as other error terms. For example, Lancaster (1979) introduces a hazard or transition rate model in which the duration depends on unobserved ability, which is denoted by a structural error term, and also depends on luck. We use an extension of this model to illustrate the concept of nonparametric identification.

A nonparametric model can have many parametric models as special cases. For example, a linear model is a parameterization of the more general model of the relationship between a regressor, an error term, and an outcome. If such a special case fails to be parametrically identified, then the nonparametric model fails to be identified as well. In other words, failure of parametric identification implies failure of nonparametric identification.

An important issue in the analysis of nonparametric identification is that some parametric special cases may be parametrically identified, even though the model is not nonparametrically identified. We illustrate this type of situation using an example. Such identified special cases may make it hard to empirically detect failures of nonparametric identification, which shows the importance of nonparametric identification proofs. Having a nonparametric identification result means that special cases can be viewed as simplified models. Without nonparametric identification, a 'simplifying assumption' may actually be an 'identifying assumption'.

While sometimes a special case of a model is parametrically identified even though the more general model is not nonparametrically identified, in other situations, both parametric identification and nonparametric identification fail. We illustrate this case with another

*We are grateful to Carter Hill for his enthusiasm and insights in econometrics and teaching econometrics. We wrote this paper for the volume of Advances in Econometrics in his honor. We hope that it will be helpful to teach Nonparametric Identification. We thank Tu Li for helpful comments and help with the numerical illustrations. This work was funded by the National Science Foundation (#2117324) and the University of Arizona,

example. An excellent explanation and review of nonparametric identification is given by Matzkin (2007).

2 Failure of Nonparametric Identification

In this section, we use an example in which we construct several parameterizations of the same nonparametric model. A nonparametric model can have many parametric models as special cases. If such a special case fails to be parametrically identified, then the nonparametric model fails to be identified as well. In our example, three special cases of the nonparametric model are parametrically identified, but a fourth case is not parametrically identified. Because parametric identification fails in at least one instance, the nonparametric model fails to be identified.

For our example, we extend the model in Lancaster (1979). Lancaster (1979) introduces a hazard model in which the duration depends on unobserved ability, which is denoted by a structural error term, and also depends on luck. The motivation to have structural error terms is to make the model more realistic and therefore more interesting. Thus, the reason to add this unobserved heterogeneity to the model is to account for ability (or other unobserved factors). Without this structural error term, the unemployment duration would depend only on luck.

We now describe our nonparametric model. Let T denote the duration, and let v denote the unobserved ability. Consider the following hazard model,

$$\theta(t|v) = v\lambda(t),$$

where $v \sim Gamma(\gamma, \delta)$ and $\gamma, \delta > 0$ and $\lambda(t)$ is the baseline hazard function that depends on time t. The conditional survival function is given by

$$\bar{F}(t|v) = \exp\{-v\Lambda(t)\},$$

where $\Lambda(t) = \int_0^t \lambda(s)ds$ is the integrated baseline hazard function. Further, the conditional survivor function equals one minus the conditional cumulative distribution function. Researchers can only estimate densities and survivor functions that do not condition on unobserved error terms. Therefore, we integrate out v. This gives an unconditional survival function,

$$\begin{aligned}
\bar{F}(t) =& \int_0^t \frac{\exp\{-v\Lambda(t)\} \cdot \delta^\gamma v^{\gamma-1} \cdot \exp\{-\delta v\}}{\Gamma(\gamma)} \\
=& \frac{\delta^\gamma}{\{\delta + \Lambda(t)\}^\gamma} \int_0^t \frac{(\Lambda(t) + \delta)^\gamma \cdot v^{\gamma-1} \cdot \exp\{-v(\delta + \Lambda(t))\}}{\Gamma(\gamma)} dv \\
=& \frac{\delta^\gamma}{\{\delta + \Lambda(t)\}^\gamma}.
\end{aligned}$$

The above unconditional survival function gives the following density,

$$f(t) = \frac{\gamma\lambda(t)\delta^{\gamma}}{\{\delta + \Lambda(t)\}^{\gamma-1}}.$$

We can approximate the baseline hazard and integrated baseline hazard by the following series approximations. These are three special parameterized cases of our nonparametric model.

Model I: Linear Approximation	$\lambda(t) = 1$	$\Lambda(t) = t$
Model II: Quadratic Approximation	$\lambda(t) = 1 + \pi_1 t$	$\Lambda(t) = t + \pi_1 t^2$
Model III: Cubic Approximation	$\lambda(t) = 1 + \pi_1 t + \pi_2 t^2$	$\Lambda(t) = t + \pi_1 t^2 + \pi_2 t^3$

The log likelihood for these models is

$$L(\alpha|T_1, ..., T_N) = \ln\gamma + \frac{1}{N}\sum_i \ln\{\lambda(t)\} + \gamma\ln\delta - (\gamma+1)\frac{1}{N}\sum_i \ln\{\delta + \Lambda(T_i))\},$$

where the parameter vector α contains all parameters, N is the sample size, and $i = 1, ..., N$. Models I through III use different series approximations to the integrated baseline hazard $\Lambda(t)$, i.e., they are different parameterizations of our nonparametric model. In our simulation, T_i, $i = 1, ..., N$, is exponentially distributed with mean 1, and these realizations are independent of each other. We estimate γ and δ and the other parameters. The mean of the Gamma distribution is $\mu = \frac{\gamma}{\delta}$, and the variance is $\sigma^2 = \frac{\gamma}{\delta^2}$.

In Tables 1, 2 and 3, we report the mean and variance of the Gamma distribution. Table 1 corresponds to Model I above, a linear approximation, Table 2 corresponds to Model II, a quadratic approximation, and Table 3 corresponds to Model III, a cubic approximation. We use 10,000 replications to calculate the means and the variances.

Models I through III are all special cases of the model with a nonparametric integrated baseline hazard $\Lambda(t)$. All three models are parametrically identified. However, not all the parameters in Models II and III are precisely estimated. We report the variations[1] in Tables 1 through 3.

The simulation results in Tables 1, 2 and 3 show that allowing for duration dependence (i.e. a non-constant hazard) substantially changes the estimated variation of the Gamma distribution. In particular, allowing for duration dependence increases the variation of the unobserved heterogeneity distribution. The variance, σ^2, is much smaller in the linear approximation than in the cubic approximation (0.168 versus 11.446 for $N = 800$). Further, Figure 1 shows the distribution of the parameter δ for each series approximation (linear, quadratic, and cubic) for $N = 3200$. Both the distribution and the mean of δ vary across the three series approximations.

[1]Data generating process: Let $U_1, ..., U_N$ be independent draws from the uniform distribution with support [0,1]. Then $T_i = -\ln(U_i)$ for $i = 1, ..., N$, are independent exponentially distributed with mean one.

Table 1: Model I (Linear Approximation) Mean and Variance

	$N = 200$	$N = 800$	$N = 3200$
$\mu = \gamma/\delta$	1.023	1.014	1.005
	(0.032)	(0.015)	(0.004)
$\sigma^2 = \gamma/\delta^2$	0.188	0.168	0.149
	(0.012)	(0.007)	(0.001)

Standard errors in parentheses

Table 2: Model II (Quadratic Approximation) Mean and Variance

	$N = 200$	$N = 800$	$N = 3200$
$\mu = \gamma/\delta$	0.967	0.989	0.996
	(0.016)	(0.001)	(0.000)
$\sigma^2 = \gamma/\delta^2$	0.592	0.354	0.318
	(1.363)	(0.000)	(0.003)
π_1	0.161	0.076	0.043
	(0.031)	(0.004)	(0.000)

Standard errors in parentheses

Table 3: Model III (Cubic Approximation) Mean and Variance

	$N = 200$	$N = 800$	$N = 3200$
$\mu = \gamma/\delta$	0.981	1.059	1.014
	(2.277)	(0.051)	(0.000)
$\sigma^2 = \gamma/\delta^2$	361.889	11.446	0.787
	(20,627.429)	(14.993)	(0.001)
π_1	0.041	0.079	0.091
	(0.026)	(0.004)	(0.001)
π_2	0.147	0.054	0.027
	(0.019)	(0.003)	(0.000)

Standard errors in parentheses

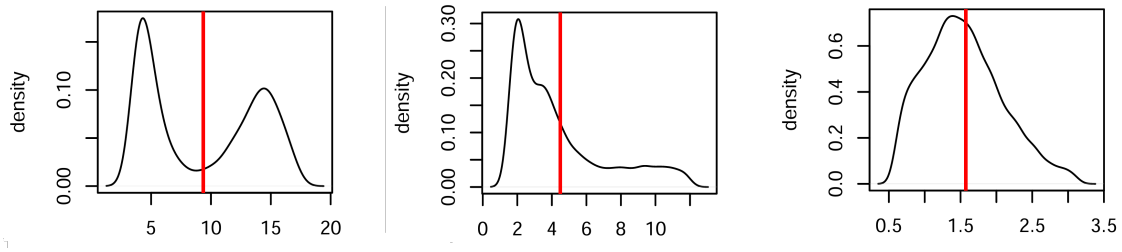

Figure 1: Distribution of δ for each series approximation for $N = 3200$

The reason for the instability of the estimate of σ^2, shown across Tables 1, 2 and 3, is that these three series approximations give different approximations to the Gompertz hazard function, and the Gompertz hazard function fails to be identified when the unobserved heterogeneity has a Gamma distribution with $\delta = 1$. The Gompertz hazard model is a fourth parametrization of our nonparametric model. However, unlike our other three parametrizations of our nonparametric model, the Gompertz hazard model is not parametrically identified.

The Gompertz hazard function is given by $\Lambda_G(t) = \exp(-\eta t) - 1$, $\lambda_G(t) = \eta \exp\{-\eta t\}$, where $\eta > 0$. Specifically,

$$\theta(t|v) = v\eta \exp(-\eta t), \eta > 0.$$

This gives the survival function

$$\bar{F}(t|v) = \exp[-v\{\exp(-\eta t) - 1\}],$$

where $v \sim Gamma(\gamma, \delta)$. Further, the unconditional survivor function for $\delta = 1$ is

$$\bar{F}(t) = \int_0^\infty \exp[-v\{\exp(-\eta t) - 1\}] \cdot \frac{\exp\{-v\} v^{\gamma-1}}{\Gamma(\gamma)} dv.$$

An identification problem arises above because $\exp[-v\{\exp(-\eta t) - 1\}]$ can be written as $\exp\{-v \exp(-\eta t)\} \cdot \exp v$. This $\exp v$ cancels with the $\exp\{-v\}$ in the second term of the above equation. This cancellation yields

$$\begin{aligned}
\bar{F}(t) =& \int_0^\infty \frac{v^{\gamma-1} \cdot \exp\{-v \exp\{-\eta t\}\}}{\Gamma(\gamma)} dv \\
=& \frac{1}{(\exp\{\eta t\})^\gamma} \int_0^\infty \frac{\exp\{\eta\gamma t\} \cdot v^{\gamma-1} \cdot \exp\{-v \exp\{-\eta t\}\}}{\Gamma(\gamma)} dv \\
=& \frac{1}{\exp\{\eta\gamma t\}}.
\end{aligned}$$

The density of T is then given by

$$f(t) = \eta\gamma \exp\{-\eta\gamma\}.$$

Thus, T has an exponential distribution with mean $\frac{1}{\eta\gamma}$. The mean of T is identified, and so is the product $\eta\gamma$. However, neither η nor γ are separately identified. Because γ is not identified, the duration dependence and the variation of the unobserved error term are not identified in the Gompertz model.

The simulation results in Tables 1 through 3 are easier to interpret when we realize this failure of identification. In particular, the simulation results show that the standard error on σ^2 is large for Model II for $N = 200$. For Model III, the standard error on σ^2 is large for $N = 200$ and $N = 800$. Models I through III are all parametrically identified, but these large standard errors show that Model II is not empirically identified for $N = 200$, and Model III is not empirically identified for $N = 200$ and $N = 800$. The reason for these large standard errors is that Model II and III approximate the Gompertz hazard model, which is not parametrically identified.

Models I through III are parametrically identified, but the Gompertz hazard model, which is also a parametrization of our nonparametric model, is not. Thus, an interpretation of the simulations is that the different series estimators yield different approximations to the Gompertz baseline and integrated baseline hazards. While all three models are parametrically identified, Models II and III are close to the Gompertz hazard model, and therefore they need a large sample size to identify the parameters in practice. $N = 800$ is not a large enough sample size to empirically identify the parameters of Model III.

3 Conclusion

Empirical researchers often want to specify a model with structural unobserved heterogeneity, as well as other error terms. The motivation to have structural heterogeneity is to make the model more realistic and therefore more interesting. For example, Lancaster (1979) introduces a hazard or transition rate model in which the duration depends on unobserved ability, which is denoted by a structural error term, and also depends on luck. We use an extension of this model to illustrate the concept of nonparametric identification.

A nonparametric model can have many parametric models as special cases. If such a special case fails to be parametrically identified, then the nonparametric model fails to be identified as well.

Without nonparametric identification, a 'simplifying assumption' may actually be an 'identifying assumption'. Such parametrically identified special cases may make it hard to empirically detect failures of nonparametric identification, which shows the importance of nonparametric identification proofs.

4 References

Lancaster, Tony. (1979): Econometric Methods for the Duration of Unemployment, *Econometrica*, 47, 939-956, https://doi.org/10.2307/1914140

Matzkin, Rosa (2007): Nonparametric Identification, *Handbook of Econometrics*, Volume 6, Part B, editors James J. Heckman and Edward E. Leamer: 5307-5368, https://www.sciencedirect.com/science/article/pii/S1573441207060734