Grouped Mixed Proportional Hazard Models with Spatial Dependence

Yoonseok Lee[1]

Department of Economics, Yale University

March, 2005

Abstract

In this paper we develop mixed proportional hazard models in discrete time when there is cross sectional duration dependence. To capture the cross sectional dependence, we use the lag of binary indicator weighted by the spatial weight matrix. We investigate the non-parametric specification for the baseline hazard and the unobserved heterogeneity. We use EM algorithm to estimate the duration model with unobserved heterogeneity and derive the observed information matrix for statistical inference.

Key words and phrases: spatial dependence, grouped duration, proportional hazard, unobserved heterogeneity, EM algorithm, observed information.

JEL classifications: C21, C23, C41, C51

1 Introduction

The duration model analyzes the length of a spell or the probability of leaving the initial state. This model is characterized by the hazard rate, which is the conditional probability of exiting the initial state at time r, given survival up to r. Since Cox's (1972) study, the proportional hazard (PH) model is one of the most widely used duration models. In the PH model, we define the hazard rate as

$$h(r; x, \beta) = h_0(r)\,\phi(x, \beta), \tag{1}$$

[0] **[Incomplete and very preliminary: Please do not cite or quote.]** This paper was originally written as the technical note for Carruthers, Guinnane and Lee (2005), *The Passage of the Uniform Small Loan Law*. I thank Timmothy W. Guinnane, Peter C. B. Phillips and Yuichi Kitamura for their helpful comment. All errors are mine, of course.

[1] 28 Hillhouse Ave., New Haven, CT 06511, U.S.A.; e-mail address: y.lee@yale.edu.

where r is the time elapsed since the beginning of the initial state, x is a vector of observed explanatory variables and β is a vector of parameters. The regression function $\phi(x, \beta)$ and the baseline hazard $h_0(r)$ are non-negative. In practice, however, the PH specification (1) is likely to have omitted explanatory variables because some determinants of the hazard cannot be observed. Lancaster (1979) pointed out the importance to account for such unobserved heterogeneity and proposed the mixed proportional hazard (MPH) model by multiplying a positive random variable v to (1), which is independent of x. It is well known that failure to account for unobserved heterogeneity causes to underestimate the hazard rate. Moreover, omitting unobserved heterogeneity results in the bias of β toward zero. The MPH model is then given by

$$h(r; x, \beta, v) = h_0(r)\,\phi(x, \beta)\,v,$$

where the regression function $\phi(x, \beta)$ captures the effect of observed explanatory variables, the baseline hazard $h_0(r)$ captures variation in the hazard over the spell or the time dependence, and a random variable v accounts for the omitted heterogeneity. Without loss of generality, we may put $\mathbb{E}(v) = 1$.

In this paper, we consider the situation that each individual's duration depends on the others'. To capture this feature, we introduce an exogenous spatial weight vector w and redefine the MPH model as

$$h(r; x, y_{t-1}, w, \beta, \rho, v) = h_0(r)\,\phi(x, w'y_{t-1}, \beta, \rho)\,v \tag{2}$$

for $t-1 \leq r < t$ and $t = 1, 2, \cdots$, where the regression function ϕ also depends on the weighted sum of the lag of binary indicator vector $w'y_{t-1}$. Each element of y_t is equal to unity if the duration of each person is over by the tth period. ρ is an additional parameter of the covariate $w'y_{t-1}$.

In the following section, we discuss more about the MPH model (2) and look at some technical conditions. Since the maximum likelihood estimation is known to be inappropriate for the MPH model, we instead develop the EM estimation procedure in Section 3. We also derive the explicit form of observed information of the EM estimators. All mathematical details are given in Appendix.

2 Grouped MPH model with spatial dependence

2.1 The model

Let τ_i be a positive and continuous random variable for the time to exit from a given state. We are interested in estimating the distribution of the duration τ_i for each individual $i = 1, 2, \cdots, n$. To find the distribution of τ_i, we assume a mixed proportional hazard (MPH) model such that the hazard rate of leaving the initial state is influenced by both observable and unobservable characteristics. More precisely, we let $z_{i,r}$ denote a vector of possibly time-varying observed covariates associated with the ith

member at time $r \geq 0$. We also let v_i denote an *i.i.d.* unobserved heterogeneity of individual i, which is independent of $z_{i,r}$ for all r. If we define $\mathbb{Z}_r$ as the covariates' path up through time r for all i, i.e., $\mathbb{Z}_r = \{z_{i,r'} : 0 \leq r' \leq r, \text{ for all } i\}$, then the hazard rate of τ_i at time r, conditional on $\mathbb{Z}_r$ and v_i, is given by

$$h_i(r|\mathbb{Z}_r, v_i) = h_0(r)\,\phi\left(\mathbb{Z}_r, \underline{\beta}\right) v_i \quad \text{for all } i \text{ and } r, \tag{3}$$

where $h_0(r) > 0$ is the baseline hazard at time r, $\underline{\beta}$ is a parameter vector, $\phi\left(\mathbb{Z}_r, \underline{\beta}\right) > 0$ and $v_i > 0$. Recall that the hazard function $h(r|\mathbb{Z}_r, v)$ is defined as the conditional probability of exiting the initial state at time r, given survival up to r:

$$h(r|\mathbb{Z}_r, v) = \lim_{dr \downarrow 0} \frac{\Pr(r \leq \tau < r + dr | \tau \geq r; \mathbb{Z}_r, v)}{dr}.$$

If we assume that, conditional on v and $\mathbb{Z}_r$, τ is independent and identically distributed (*i.i.d.*) with a distribution function $F_\tau(r)$ for $r > 0$, then it is well known that

$$F_\tau(r|\mathbb{Z}_r, v) = 1 - S(r|\mathbb{Z}_r, v) \quad \text{and} \quad dF_\tau(r|\mathbb{Z}_r, v)/dr = h(r|\mathbb{Z}_r, v)\, S(r|\mathbb{Z}_r, v),$$

where $S(r|\mathbb{Z}_r, v) = \Pr(\tau \geq r|\mathbb{Z}_r, v)$ is the survivor function. More details can be found in Van den Berg (2000) or in standard textbooks. Now we let t be the discrete time index such that $t = 1, 2, \cdots, K$. Like the binary choice model, we define a binary indicator $y_{i,t}$, which is equal to unity if the duration of i ends in the interval $[t-1, t)$. Thus, $y_{i,t+1} = 1$ if $y_{i,t} = 1$. Similarly, $c_{i,t}$ is a binary censoring indicator equal to one if the duration of i is censored in $[t-1, t)$. Obviously, $c_{i,K} = 1$ and $c_{i,t} = 1$ implies $c_{i,t+1} = 1$ for all i. If the duration is censored in $[t-1, t)$, i.e., $c_{i,t} = 1$, then we set $y_{i,t} = 1$ as convention.

As we specified in (2), we introduce cross sectional duration dependence in the model. More precisely, we allow that the hazard rate (3) of each individual i depends on the previous choice paths of others. For this purpose, we introduce an $n \times n$ exogenous matrix $W = (w_1, \cdots, w_n)'$, where $w_i = (w_{i1}, w_{i2}, \cdots, w_{in})'$ is an $n \times 1$ normalized spatial weight vector for each i such that $\sum_{j=1}^{n} w_{ij} = 1$. We then let

$$\phi\left(\mathbb{Z}_r, \underline{\beta}\right) = \exp\left(x'_{i,r}\beta + \rho w'_i y_{t-1}\right)$$

for $r \in [t-1, t)$, where $y_t = (y_{1t}, \cdots, y_{nt})'$, $x_{i,r}$ is a $k \times 1$ vector of covariates, β a $k \times 1$ parameter vector and ρ a scalar parameter. Conventionally[2], we define $w_{ij} = g\left[d\left(\varphi_i, \varphi_j\right)\right] \geq 0$ for $i, j = 1, 2, \cdots, n$, where $w_{ii} = 0$ and $d(\cdot, \cdot)$ is a proxy of economic distance between i and j. More precisely, $d(\cdot, \cdot)$ is a

[2] The specification of spatial weight matrix is rather arbitrary since the economic distance cannot be observed explicitly. Some examples are discussed in Appendix.

distance function of a pair of characteristics φ_i and φ_j, and $g(\cdot)$ is a nonnegative and strictly decreasing function with $g(0) = 0$. Then W is a well-defined spatial weight matrix and it controls the degree of cross sectional dependence based on the economic distance between two areal units. It captures the interactions among the economic agents and it can be rephrased as the spill-over effect or the externality. More discussions on how to specify spatial weight matrices and some examples are provided in Appendix. For more technical details, readers are referred to spatial econometrics or spatial statistics literature such as Anselin (1988), Cressie (1993), Kelejian and Prucha (1999), Lee (2002) and the references therein, to name a few. A large number of applications can also be found in regional, agricultural, environmental, and industrial organization economics literatures. See the survey by Anselin and Bera (1998) for the references.

Heckman and Singer (1984) suggest that the distribution of the unobserved heterogeneity be nonparametrically estimated in order to avoid any misspecification problem. They show that as the number of mass points increases, discrete distributions can approximate any distribution arbitrarily well. Their nonparametric estimator, however, is very sensitive to the assumed shape of the baseline hazard function. (e.g., Trussell and Richards, 1985). Meyer (1990, 1995) proposes a solution using piecewise constant baseline hazard functions of Prentice and Gloeckler (1978). By combining two approaches of Prentice and Gloeckler and of Heckman and Singer, Meyer allows (approximately) nonparametric specifications in both the unobserved heterogeneity v and the baseline hazard h_0. In our model (3), we follow this combination approach such that the baseline hazard h_0 is piecewise constant:

$$\log\left(\int_{t-1}^{t} h_0(r)\,dr\right) = \gamma_t \in \mathbb{R} \quad \text{for all } r \in [t-1, t) \text{ and } t = 1, 2, \cdots, K, \tag{4}$$

and v_i is *i.i.d.* with a discrete distribution with m supports:

$$f(v) = \sum_{j=1}^{m} p_j \mathbf{1}_{v=q_j} \tag{5}$$

for all $i = 1, 2, \cdots, n$, where $p_m = 1 - \sum_{j=1}^{m-1} p_j$, $0 < p_j < 1$ and $0 < q_j < \infty$ for all $j = 1, 2, \cdots, m$. $\mathbf{1}_A$ is the indicator function, which equals to unity if the condition A is true. The piecewise constant baseline hazard (4) is very useful especially when the hazard rate has much fluctuation or frequent peaks[3]. It extracts common deterministic time trends from the covariates as the usual fixed time effects in panel regressions. The nonparametric approach (Heckman and Singer, *op.cit.*) leads to a finite mixture model (5), where the random effect v_i is assumed to have a discrete distribution that assigns

[3] As Diebold, Rudebusch and Sichel (1993), we could also consider a pth order polynomial baseline hazard, i.e., $h_0(r) = \exp\left(\sum_{s=0}^{p} b_s t^s\right)$ for $r \in [t-1, t)$ and $t = 1, 2, \cdots, K$, where $p \to \infty$ as $K \to \infty$ and $b_s \in \mathbb{R}$ for all s.

probabilities $p_1, p_2, \cdots, p_m$ to the fixed m points of supports $q_1, q_2, \cdots, q_m$. The finite number of mass points should be set in advance to avoid an incidental parameter problem. In the theoretical point of view it is convenient to assume that the mean of v_i is one, although for the estimation purpose it is often more convenient to leave q_j $(j = 1, 2, \cdots, m)$ unrestricted and omit the constant term from the baseline hazard. In fitting the model (3), we thus simply let $\gamma_1 = 0$. We start with two points of support $(m = 2)$ and keep adding more points of support as long as all the estimated q_j are distinct. As convention, we may also normalize one location $q_m = 1$ so that we can interpret that an individual i with type $v_i = q_j$ $(j \neq m)$ is q_j-times more likely to exit the initial state than an individual with type $q_m\,(= 1)$.

Finally, note that conventional economic panel data provide observations on failure times aggregated up to discrete intervals. We therefore observe discrete duration data T_i (or grouped duration data) rather than continuous observations τ_i, where $\tau_i \in [T_i - 1, T_i)$. We can easily handle this problem[4] by splitting the time line into $K+1$ intervals: $[r_0, r_1), [r_1, r_2), \cdots, [r_{K-1}, r_K), [r_K, \infty)$ with letting $r_0 = 0$ without loss of generality, where the length of each interval corresponds to the panel survey frequency. Survival up to time r_t is thus the same as surviving until the tth interval $[r_{t-1}, r_t)$. Notice that r_K is the last time of our observation and all durations lasting over r_K are censored. Since we are usually dealing with equi-spaced panel data (i.e., observed at regular intervals such as monthly, quarterly, yearly, etc.), we can let $r_t = t$ for all $t = 1, 2, \cdots, K, K+1$ in these cases and consider $K+1$ intervals: $[0, 1), [1, 2), \cdots, [K-1, K), [K, \infty)$. Hereafter, we assume observations of failure times T_i over the discrete periods $t = 1, 2, \cdots, K$ for individuals $i = 1, 2, \cdots, n$. We further assume that covariates are at best recorded up to intervals and thus $z_{i,r} = z_{i,t}$ for all r in the tth interval $[t-1, t)$. Therefore, the hazard rate (3) of τ_i conditional on $\mathbb{Z}_t$ and v_i is actually given by

$$h_i\left(r|\mathbb{Z}_t, v_i\right) = h_0\left(r\right)\phi\left(\mathbb{Z}_t, \underline{\beta}\right) v_i$$

at time t for $r \in [t-1, t)$. However, we use the hazard rate of the discrete failure time T_i, conditional on $\mathbb{Z}_t$ and v_i, which is given by

$$h_i\left(t|\mathbb{Z}_t, v_i\right) = h_0\left(t\right)\phi\left(\mathbb{Z}_t, \underline{\beta}\right) v_i$$

for all $t = 1, 2, \cdots, K$ and $i = 1, 2, \cdots, n$.

[4] For further discussions on discretization, readers are referred to Heckman and Singer (1984), Lancaster (1990) and Sueyoshi (1995).

2.2 Technical conditions

In sum, we consider an MPH model given by

$$\begin{aligned} h_i\left(t|v_i\right) &= \exp\left(\gamma_t + x_{i,t}'\beta + \rho w_i' y_{t-1}\right) v_i \\ &= \exp\left(\gamma_t + x_{i,t}'\beta + \rho \sum_{j=1}^{n} w_{ij} y_{j,t-1}\right) v_i \end{aligned} \tag{6}$$

for $t = 1, 2, \cdots, K$ and $i = 1, 2, \cdots, n$. For expositional simplicity we suppress the dependence on the observable items x, y and W, hereafter. Recall that $x_{i,t}$ is a covariate vector containing both time-varying and time-invariant variables. $\exp\left(\gamma_t\right)$ is the piecewise constant baseline hazard. The weighted sum of $y_{i,t-1}$ by the spatial weight vector w_i can be interpreted as the average influence of other agents' past decisions on i, i.e., the cross sectional interaction. We first assume the following conditions. Here, $\mathbb{Z}_t$ still implies the complete path of the covariates $(x_{i,t}, w_i' y_{t-1})$ up to time t for all i.

Assumption FT (failure time) *(i) The failure time $T_i > 0$ is i.i.d. across i conditional on $\mathbb{Z}_t$ and v for each t; (ii) T_i is independent of $c_{j,t}$ for all i, j and t.*

Assumption UH (unobserved heterogeneity) *(i) The unobserved heterogeneity $v_i > 0$ is i.i.d. of finite mixture* (5) *with mean one; (ii) v_i is independent of $\mathbb{Z}_t$ and $c_{j,t}$ for all i, j and t.*

Assumption BH (baseline hazard) *The baseline hazard $h_0\left(t\right)$ is nonnegative for all t and piecewise constant as* (4)*.*

Heckman and Singer (1984) assume that the distribution of the censoring variable $c_{i,t}$ is known and independent of the covariate $(x_{i,t}, w_i' y_{t-1})$ to show the consistency of maximum likelihood estimators with nonparametric unobserved heterogeneity v_i. In our case, the censoring only occurs at the fixed time K when the panel survey is over, so those assumptions hold automatically. Note that v_i is only observed by individual i, not by the econometrician. Its mean is usually normalized to one so that the expected hazard rate is the unconditional hazard rate, *viz.*,

$$\mathbb{E}\left[h_i\left(t|v_i\right)|\mathbb{Z}_t\right] = h_0\left(t\right)\exp\left(x_{i,t}'\beta + \rho w_i' y_{t-1}\right)\mathbb{E}\left(v_i\right) = h_i\left(t\right),$$

where $h_i\left(t\right) = h_0\left(t\right)\exp\left(x_{i,t}'\beta + \rho w_i' y_{t-1}\right)$. Recall that v_i is independent of $\mathbb{Z}_t$. (Assumption UH-(ii)) We now assume further condition for the identification of h_0, β, ρ and the distribution of v.

Assumption CV (covariates) *(i)* $(x_{i,t}, w_i'y_{t-1}) \in \mathcal{X} \times \mathcal{Y}$*, where* $\mathcal{X}$ and $\mathcal{Y}$ *are open sets in* $\mathbb{R}^k$ *and* $\mathbb{R}$ *respectively. (ii) At least one argument of* $x_{i,t}$ *is defined on the continuum. (iii) The regression function* $\exp\left(x_{i,t}'\beta + \rho w_i'y_{t-1}\right)$ *is nonconstant on* $\mathcal{X} \times \mathcal{Y}$*.*

As shown in Elbers and Ridder (1982), Assumption (UH), (BH) and (CV) yield identifiability of our model (6) at least up to constant. Notice that Assumption (UH) is the same as Assumption 1 of Elbers and Ridder (*op.cit.*). Assumption (BH) satisfies Assumption 2 (*ibid.*) because $\int_0^t h_0(s)\,ds = \sum_{s=1}^t \exp(\gamma_s)$ is an increasing function of $t \geq 0$. Assumption (CV) and the form of the proportional hazard function satisfy Assumption 3 (*ibid.*) because the regression function $\exp\left(x_{i,t}'\beta + \rho w_i'y_{t-1}\right)$ is obviously differentiable with respect to $\left(\beta', \rho\right)'$. Further discussions of identification with unobserved heterogeneity can be found in Kiefer and Wolfowitz (1956), Heckman and Singer (1984) or Meyer (1995). These studies look at the very general situation that the number of supports of v increases to infinity. Finally, the following condition for the spatial weight matrix $W = (w_{ij})$ is also important.

Assumption SW (spatial weight matrix) *(i) The spatial weight matrix* W *is predetermined and correctly specified; (ii) each element of* W *is nonnegative and all the diagonal elements of* W *are zero; (iii)* W *is row normalized, i.e.,* $\sum_{j=1}^n w_{ij} = 1$ *for all* i*; (iv)* W *is independent of* v_i *for all* i*, and* $w_{ij} \notin \mathcal{X}$ *for all* i *and* j*.*

It is important to assume that the spatial weight matrix W is predetermined (Assumption SW-(i)) and independent of v_i (Assumption SW-(iv)). Pre-specifying W outside the model is an easy way to obtain a predetermined and exogenous W. It prevents any identification problem as pointed by Manski (1993). The independence assumption also guarantees the absence of endogeneity problems since all the diagonal elements of W are zero by construction and Assumption (UH) holds. Keeping the spatial weight matrix W out of the covariate space $\mathcal{X}$ prevents any possible multicollinearity problem[5] between $x_{i,t}$ and $w_i'y_{t-1}$. Therefore, exogeneity of W is crucial for the satisfactory properties of the estimators $\widehat{\beta}$ and $\widehat{\rho}$.

When we consider the asymptotics for large cross sectional size n, different from time series, we can think of two distinct scenarios: the in-filling asymptotics (to increase the number of observations within a fixed boundary) and the increasing-domain asymptotics (to expand the boundary of samples as time series) in spatial econometrics. In the applied spatial models, however, each cross sectional unit i has a limit number of economic neighbors regardless of the sample size n. Therefore, when we work with the increasing-domain asymptotics, we are expecting that the spatial weight matrix W becomes sparser as n grows. In the in-filling asymptotics, on the other hand, we cannot expect W becomes sparser as n

[5] If we use any variables $\varphi_{i,t}$ that are included among $x_{i,t}$ to construct the spatial weight matrix W and if W happens to be a linear transformation of $\varphi_{i,t}$, then $x_{i,t}$ becomes collinear with the spatial weight vector w_i and so does with $w_i'y_{t-1}$.

grows. In this case, the row normalization (Assumption SW-(iii)) is important to prevent the weighted sum $w_i' y_{t-1}$ from exploding. It controls the degree of cross sectional dependence so that the M-estimator has proper asymptotic properties. More discussions on the asymptotic properties can be found in Sargan (1975) in terms of a large system. Kelejian and Prucha (1999) and Lee (2002, 2004) provide more general and technical conditions in the spatial econometrics context.

3 Estimation using EM algorithm

3.1 Log-likelihood function

Let us first consider the ith individual, who drops out of the sample in the T_ith interval $[T_i - 1, T_i)$ either by exiting the initial state ($c_{i,T_i} = 0$) or by censoring ($c_{i,T_i} = 1$). The conditional log-likelihood function on the unobserved v is given by

$$\log L_{(t \leq T_i)|(t=0)}(\theta|v) = \sum_{s=1}^{T_i} \log L_{(t=s)|(t \leq s-1)}(\theta|v),$$

where θ is the parameter vector and the (joint) log-likelihood is conditional on the initial state at $t = 0$. If we ignore the conditioning on the initial state, the conditional log-likelihood function is then given by

$$\log L(\theta|v) = \sum_{i=1}^{n} \left\{ \delta_i \log\left(1 - \exp\left(-\exp\left(D_{i,T_i}'\theta\right)\right) v_i\right) - \sum_{s=1}^{T_i - 1} \exp\left(D_{i,s}'\theta\right) v_i \right\} \tag{7}$$

using the specification (6) similarly as Prentice and Gloeckler (1978) and Meyer (1990, 1995), where $\delta_i = (1 - c_{i,T_i})$, $\theta = \left(\gamma_1, \gamma_2, \cdots, \gamma_K, \beta', \rho\right)'$, and the covariates vector $D_{i,t} = \left(e_t', x_{i,t}', w_i' y_{t-1}\right)'$ with e_t being the tth column of the identity matrix I_K. Integrating (7) over the distribution (5) of v yields the unconditional log-likelihood:

$$\log L(\theta) = \sum_{j=1}^{m} p_j \log L(\theta|v = q_j). \tag{8}$$

As noted in Lee (2000), Lancaster (1990, Chapter 8.4) and Heckman and Singer (1984), however, the maximum likelihood (ML) estimation is not appropriate on the mixture model (8). It is because the parameters of the heterogeneity distribution are not guaranteed to lie on the interior of a compact set. Moreover, numbers of studies report that the ML estimation of mixture models has convergence problem when the models have both the piecewise constant baseline hazard and the finite mixture unobserved heterogeneity.

Instead, the appropriate choice of algorithm for fitting a mixture model is the EM (Expectation-

Maximization) algorithm (Dempster, Laird and Rubin, 1977). The EM algorithm was originally invented to cope with inference in models imposing missing data. In our case, the unobserved heterogeneity v is essentially a problem of missing data. To fix this idea, we now consider an alternative expression of the conditional log-likelihood function (7) for person i, which is given by

$$\begin{aligned}\log L_i\left(\theta|v_i=q_j\right) &= \sum_{j=1}^{m}\lambda_{ij}\left\{\delta_i\log\left(1-\exp\left(-\exp\left(D'_{i,T_i}\theta\right)q_j\right)\right)-\sum_{s=1}^{T_i-1}\exp\left(D'_{i,s}\theta\right)q_j\right\}\\ &= \sum_{j=1}^{m}\lambda_{ij}\log L_i^*\left(\theta,q_j\right),\end{aligned}$$

where

$$\log L_i^*\left(\theta,q_j\right)=\delta_i\log\left(1-\exp\left(-\exp\left(D'_{i,T_i}\theta\right)q_j\right)\right)-\sum_{s=1}^{T_i-1}\exp\left(D'_{i,s}\theta\right)q_j.$$

The unobservable (or missing) data λ_{ij} is a binary indicator equal to unity if the value of v_i is q_j and zero otherwise. For the *prior* probabilities p_j $(j=1,2,\cdots,m)$, the log of the density function of v_i is

$$\log f\left(v_i\right)=\sum_{j=1}^{m}\lambda_{ij}\log p_j.$$

from (5). Hence the unconditional joint log-likelihood function of both the observed and the unobserved items is defined as

$$\begin{aligned}\log L\left(\Theta\right) &= \sum_{i=1}^{n}\left\{\log f\left(v_i\right)+\log L_i\left(\theta|v_i\right)\right\}\\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_{ij}\left\{\log p_j+\log L_i^*\left(\theta,q_j\right)\right\},\end{aligned}\tag{9}$$

where $\Theta=\left(\theta',\mathbf{p}',\mathbf{q}'\right)'$, $\mathbf{p}=\left(p_1,p_2,\cdots,p_{m-1}\right)'$ and $\mathbf{q}=\left(q_1,q_2,\cdots,q_m\right)'$. Note that[6] p_m is automatically determined from the restriction of probabilities, $p_m=1-\sum_{j=1}^{m-1}p_j$. As noted in Guo and Rodriguez (1992), the M-step of the EM algorithm only requires numerical maximization of $\log L_i^*\left(\theta,q_j\right)$.

3.2 EM algorithm

The EM algorithm proceeds in two steps. The E-step calculates the conditional expectation of λ_{ij} given observed data, δ_i and $\mathbb{D}_i=\left\{D_{i,s}:0\leq s\leq T_i\right\}$, and given the likelihood L_i^* evaluated at the current

[6] In this formula, all the elements of $\mathbf{q}$ and the parameters $(\gamma_1,\gamma_2,\cdots,\gamma_T)$ of the piecewise constant baseline hazard have no restriction. As noted in the previous chapter, however, proper normalization is necessary in practice and we usually let $\gamma_1=0$ and $q_m=1$.

parameter estimates. It can be shown that the *posterior* probability of $v_i = q_j$ is derived as

$$\mathbf{E}\left(\lambda_{ij}|\delta_i, \mathbb{D}_i\right) = \frac{p_j L_i^*\left(\theta, q_j\right)}{\sum_{\ell=1}^{m} p_\ell L_i^*\left(\theta, q_\ell\right)} \equiv \pi_{ij} \quad \text{for all } i \text{ and } j. \tag{10}$$

If we let $\widehat{\pi}_{ij}$ denote π_{ij} evaluated at the current parameter estimates, substituting $\widehat{\pi}_{ij}$ for λ_{ij} in (9) gives the outcome of the E-step:

$$Q\left(\Theta\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{\pi}_{ij} \log p_j + \sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{\pi}_{ij} \log L_i^*\left(\theta, q_j\right). \tag{11}$$

Then, the M-step consists of maximizing $Q\left(\Theta\right)$ with respect to $\Theta = \left(\theta', \mathbf{p}', \mathbf{q}'\right)'$. Notice that the formula (11) implies we can split the maximization problem for $(\theta, \mathbf{q}, \mathbf{p})$ into two sub-optimization problems for $\mathbf{p}$ and $(\theta, \mathbf{q})$ respectively. Maximization with respect to p_j's gives an explicit solution

$$\widehat{p_j} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\pi}_{ij}$$

from a general result in finite mixture models. (Everitt and Hand, 1981) Maximization with respect to $(\theta, \mathbf{q})$ can be done with conventional optimization procedures. We iterate these whole E and M-steps until the estimates converge.

We have several remarks on running the EM algorithm. The initial values for the EM algorithm can be chosen from the ML estimates of the duration model without unobserved heterogeneity, which is given by

$$h_i\left(t\right) = \exp\left(\gamma_t + x_{i,t}'\beta + \rho w_i' y_{t-1}\right).$$

We then start the EM algorithm on the duration model with unobserved heterogeneity:

$$h_i\left(t|u_i\right) = \exp\left(\gamma_t + x_{i,t}'\beta + \rho w_i' y_{t-1} + u_i\right)$$

using the first step ML estimates $\left(\gamma_1^0, \cdots, \gamma_T^0, \beta^0, \rho^0\right)$ and arbitrary $\left(\mathbf{q}^0, \mathbf{p}^0\right)$ as the initial values. Notice that we reparametrize the model (6) in that $v_i = \exp(u_i)$. We then do not need any restrictions on u_i because $0 < v_i < \infty$ is satisfied for any $-\infty < u_i < \infty$. As we discussed, we start with two points of support (q_1, q_2) for the unobserved heterogeneity v and keep adding more points of support as long as all the estimates $\widehat{q}_j$ are distinct. We normalize one location $q_1 = 1$ but we leave the mean of v unrestricted by omitting the first term of baseline hazard, i.e., $\gamma_1 = 0$. We may also conduct the Simulated Annealing (Goffe et al., 1994; or Goffe, 1996) grid search to find and/or confirm the global maximum.

3.3 Observed information matrix

The EM algorithm is known to be robust to the choice of initial values and practically guarantees convergence to at least a local maximum. One of its disadvantages[7] is that it dose not provide standard errors as an immediate by-product unlike the Newton-Raphson type methods. Therefore, even though the estimation with unobserved heterogeneity is very common, the literature is rarely informative about the precision of the estimators. This happens more frequently in case of the MPH model with discrete mixture unobserved heterogeneity.

Louis (1982), Guo and Rodriguez (1992) and Oakes (1999), however, show how to find the observed information matrix within the EM algorithm framework. We then can easily obtain the asymptotic variance matrix estimate from the inverse[8] of the observed information matrix. More precisely, under the regularity conditions, Louis (1982) shows that the observed information matrix $\mathcal{I}$ can be obtained by

$$\mathcal{I}(\Theta) = \mathbb{E}_v\left(\frac{\partial^2}{\partial\Theta\partial\Theta'}\log L(\Theta)\right) - \mathbb{V}_v\left(\frac{\partial}{\partial\Theta}\log L(\Theta)\right),$$

where $L(\Theta)$ is the complete data likelihood function in (9). The expectation $\mathbb{E}_v(\cdot)$ and variance $\mathbb{V}_v(\cdot)$ are taken over the conditional distribution of v given the observed data $\{\delta_i, \mathbb{D}_i\}_{i=1}^n$. As noted in Guo and Rodriguez (1992), the first term can be interpreted as the conditional expectation of the observed information when v is observed. The second term represents the missing information associated with the conditional distribution of v given the observed data. We derive the explicit form of the observed information matrix $\mathcal{I}$ in Appendix.

Appendix

A.1 Spatial weight matrix

In this section, we introduce the spatial weight matrix and discuss how to specify it. We let φ_i be a cross-sectional economic variable[9] for $i = 1, \cdots, n$. For each location (or individual) indexed by i, we consider a function $w_{ij} : \mathbb{R}^2 \to \mathbb{R}_+$ such that $w_{ij} = g\left[d\left(\varphi_i, \varphi_j\right)\right] \geq 0$ for all $j = 1, 2, \cdots, n$. $d(\cdot,\cdot)$ is a (proxy of) economic distance between i and j and it is a metric (or a distance function) of a pair of cross sectional characteristics φ_i and φ_j. $g(\cdot)$ is a nonnegative and strictly decreasing function with $g(0) = 0$. We

[7] Another major disadvantage is its slow convergence, which is closely linked to the asymptotic variance matrix problem. There are several studies on the modification of the EM algorithm to speed up the convergence, such as Louis (1982), Meng and Rubin (1991), and Oakes (1999) to name a few.

[8] Or we can use the generalized inverse if necessary.

[9] For the simplicity, we only consider a time-invariant spatial weight matrix. Time-varying generalization is straightforward if we use time-varying cross-section data, i.e., panel data.

normalize w_{ij} as $\sum_{j=1}^{n} w_{ij} = 1$ and let $w_{ii} = 0$ for all i so that an $n \times 1$ vector $w_i = (w_{i1}, w_{i2}, \cdots, w_{in})'$ becomes a spatial *weight* vector for each i.

As examples[10] of a spatial weight matrix $W = (w_1, \cdots, , w_n)'$, we consider the U.S. state-wise data. In this example, therefore, $i\,(= 1, 2, \cdots, n)$ is the index of each state. We can think of the following variables that determine the spatial weight matrix W. **(i)** The geographical location such as whether two different states are neighbors, and whether the state belongs to the northern states or not. **(ii)** The measurable economic distance such as the distance between two capitals and the transportation cost. **(iii)** The similarity of demographic characteristics such as population, labor structure, race distribution, income level, and the leading political party. **(iv)** The influence of major states measured by the size of industry or market.

We then can determine the economic spatial weight matrix for each case as follows. For variables **(i)** $w_{ij} = b_{ij}/B_i$, where b_{ij} is an indicator equal to unity if state i and j are neighbors, and B_i is the total number of neighbors of state i. For variables **(ii)** $w_{ij} = d_{ij}^{-1} / \sum_{k \neq i} d_{ik}^{-1}$, where d_{ij} is the geographical distance or transportation cost between two different states i and j. For variables **(iii-a)** $w_{ij} = |e_i - e_j|^{-1} / \sum_{k \neq i} |e_i - e_k|^{-1}$, where $e_i = (1/K) \sum_{t=1}^{K} e_{i,t}$ is the average level of the time-varying demographic characteristic (or the time-invariant characteristic) of state i. Notice that the more similar are e_i and e_j, the larger is $|e_i - e_j|^{-1}$. If we directly use $e_{i,t}$ instead of e_i, then we get the time-varying spatial weight matrix in this case. For variables **(iii-b)** $w_{ij} = 1_{ij} / \sum_{k \neq i} 1_{ik}$, where 1_{ij} is an indicator equal to unity if state i and j share a common binary characteristic. For variables **(iv)** $w_{ij} = |sz_j| / \sum_{k \neq i} |sz_k|$, where sz_i is the measure of the market spill-over power of state i. It could be the number of workers or the market concentration ratio. The size of manufacturing sector is also a good example. More examples on constructing spatial weight matrices can be found in Case and Rosen (1993), Pinske and Slade (1998), and Pinske, Slade and Brett (2002).

Note that if we obtain more than one weight matrix, i.e., ℓ spatial sub-weight matrices $W_1, W_2, \cdots, W_\ell$, we then can formulate the general convex combination as

$$W = \mu_1 W_1 + \mu_2 W_2 + \cdots + \mu_\ell W_\ell,$$

where μ_j are any constants satisfying $\sum_{j=1}^{\ell} \mu_j = 1$. Notice that the spatial weight matrix W still satisfies the normalization assumption such that $\sum_{j=1}^{n} w_{ij} = 1$ and $w_{ii} = 0$ for all i. μ_j can be pre-specified

[10] As another example, we can parametrize the spatial weight matrix as $W(\eta) = (w_{ij}(\eta))_{i,j=1}^{n}$ such that $w_{ij}(\eta) = \exp\left(-d\left(\varphi_i, \varphi_j\right)/\eta\right)$. In this case, we need to estimate η along with other parameters, which determines the rate that spatial influence attenuates with distance.

based on the importance of each economic variable, or they can be estimated in a properly modified model. A straightforward example is a convex combination of two different spatial weight matrices W_1 and W_2 such as $W = \mu W_1 + (1-\mu) W_2$, and we can estimate μ (instead of ρ) in our model.

A.2 Observed information matrix

Louis (1982) shows that the observed information matrix $\mathcal{I}$ can be obtained from

$$\mathcal{I}(\Theta) = \mathbb{E}_v\left(\frac{\partial^2}{\partial\Theta\partial\Theta'}\log L(\Theta)\right) - \mathbb{V}_v\left(\frac{\partial}{\partial\Theta}\log L(\Theta)\right) \equiv \mathcal{I}_1(\Theta) - \mathcal{I}_2(\Theta).$$

We will investigate these two terms separately.

Complete data information matrix $\mathcal{I}_1(\Theta)$: Under the regularity condition, the first term (the complete data information matrix) is equal to $-\partial^2 Q(\Theta)/\partial\Theta\partial\Theta'$ and thus it can be obtained as a by-product of the E-step. From the formula (11) and for person i,

$$Q_i(\Theta) = \sum_{j=1}^{m}\widehat{\pi}_{ij}\log p_j + \sum_{j=1}^{m}\widehat{\pi}_{ij}\left\{\delta_i\log\left(1-\exp\left(-q_j\alpha_{i,T_i}\right)\right) - \sum_{s=1}^{T_i-1}q_j\alpha_{i,s}\right\},$$

where $\alpha_{i,t} = \exp\left(D_{i,t}'\theta\right)$. Since the maximization procedures of $\mathbf{p}$ and $(\theta,\mathbf{q})$ can be separated, the Hessian matrix of $Q_i(\Theta)$ is block diagonal. The Hessian matrix corresponding to $(\theta,\mathbf{q})$ is automatically obtained from the M-step. More precisely,

$$\begin{aligned}
\frac{\partial}{\partial\theta}Q_i(\Theta) &= \sum_{j=1}^{m}\widehat{\pi}_{ij}\left\{\delta_i\frac{q_j\alpha_{i,T_i}\exp\left(-q_j\alpha_{i,T_i}\right)}{1-\exp\left(-q_j\alpha_{i,T_i}\right)}D_{i,T_i} - \sum_{s=1}^{T_i-1}q_j\alpha_{i,s}D_{i,s}\right\};\\
\frac{\partial}{\partial q_k}Q_i(\Theta) &= \widehat{\pi}_{ik}\left\{\delta_i\frac{\alpha_{i,T_i}\exp\left(-q_k\alpha_{i,T_i}\right)}{1-\exp\left(-q_k\alpha_{i,T_i}\right)} - \sum_{s=1}^{T_i-1}\alpha_{i,s}\right\} \quad \text{for } k=1,2,\cdots,m;\\
\frac{\partial}{\partial p_k}Q_i(\Theta) &= \frac{\widehat{\pi}_{ik}}{p_k} - \frac{\widehat{\pi}_{im}}{p_m} \quad \text{for } k=1,2,\cdots,m-1,
\end{aligned}$$

where $p_m = 1-\sum_{j=1}^{m-1}p_j$. If we denote

$$A_{ij} = \frac{\alpha_{i,T_i}\exp\left(-q_j\alpha_{i,T_i}\right)}{1-\exp\left(-q_j\alpha_{i,T_i}\right)}, \tag{A1}$$

then

$$\frac{\partial^2 Q_i(\Theta)}{\partial\theta\partial\theta'} = \sum_{j=1}^{m}\widehat{\pi}_{ij}\left\{\delta_i\left(q_jA_{ij}-q_j^2\alpha_{i,T_i}A_{ij}-q_j^2A_{ij}^2\right)D_{i,T_i}D_{i,T_i}'-\sum_{s=1}^{T_i-1}q_j\alpha_{i,s}D_{i,s}D_{i,s}'\right\}; \quad \text{(A2)}$$

$$\frac{\partial^2 Q_i(\Theta)}{\partial\theta\partial q_k} = \widehat{\pi}_{ik}\left\{\delta_i\left(A_{ik}-\alpha_{i,T_i}A_{ik}-q_kA_{ik}^2\right)D_{i,T_i}-\sum_{s=1}^{T_i-1}\alpha_{i,s}D_{i,s}\right\} \quad \text{for } k=1,2,\cdots,m; \quad \text{(A3)}$$

$$\frac{\partial^2 Q_i(\Theta)}{\partial q_k\partial q_\ell} = -\widehat{\pi}_{ik}\delta_i\left(\alpha_{i,T_i}A_{ik}+A_{ik}^2\right)\mathbf{1}_{k=\ell} \quad \text{for } k,\ell=1,2,\cdots,m; \quad \text{(A4)}$$

$$\frac{\partial^2 Q_i(\Theta)}{\partial p_k\partial p_\ell} = -\frac{\widehat{\pi}_{ik}}{p_k^2}\mathbf{1}_{k=\ell}+\frac{\widehat{\pi}_{im}}{p_m^2} \quad \text{for } k=1,2,\cdots,m-1. \quad \text{(A5)}$$

Therefore, the complete data information matrix is given by the negative sum of individual Hessians:

$$\mathcal{I}_1(\Theta) = -\sum_{i=1}^{n}\frac{\partial^2}{\partial\Theta\partial\Theta'}Q_i(\Theta),$$

where $\partial^2 Q_i(\Theta)/\partial\Theta\partial\Theta'$ consists of the second derivatives (A2) to (A5) conformably with the array of parameters $\Theta=\left(\theta',q_1,\cdots,q_m,p_1,\cdots,p_{m-1}\right)'$. All other terms are zero since $\partial^2 Q_i(\Theta)/\partial\Theta\partial\Theta'$ is block diagonal.

Missing information matrix $\mathcal{I}_2(\Theta)$: From the formula (9) and for person i,

$$\log L_i(\Theta) = \sum_{j=1}^{m}\lambda_{ij}\log p_j+\sum_{j=1}^{m}\lambda_{ij}\left\{\delta_i\log\left(1-\exp\left(-q_j\alpha_{i,T_i}\right)\right)-\sum_{s=1}^{T_i-1}q_j\alpha_{i,s}\right\},$$

and

$$\frac{\partial}{\partial\theta}\log L_i(\Theta) = \sum_{j=1}^{m}\lambda_{ij}\left\{\delta_iq_jA_{ij}D_{i,T_i}-\sum_{s=1}^{T_i-1}q_j\alpha_{i,s}D_{i,s}\right\}\equiv\sum_{j=1}^{m}\lambda_{ij}B_{ij};$$

$$\frac{\partial}{\partial q_k}\log L_i(\Theta) = \lambda_{ik}\left\{\delta_iA_{ik}-\sum_{s=1}^{T_i-1}\alpha_{i,s}\right\}\equiv\lambda_{ik}C_{ik} \quad \text{for } k=1,2,\cdots,m;$$

$$\frac{\partial}{\partial p_k}\log L_i(\Theta) = \frac{\lambda_{ik}}{p_k}-\frac{\lambda_{im}}{p_m} \quad \text{for } k=1,2,\cdots,m-1,$$

where A_{ij} is defined as $(A1)$ and

$$B_{ij} = \delta_iq_jA_{ij}D_{i,T_i}-\sum_{s=1}^{T_i-1}q_j\alpha_{i,s}D_{i,s},$$

$$C_{ij} = \delta_iA_{ij}-\sum_{s=1}^{T_i-1}\alpha_{i,s}.$$

Therefore,

$$\mathbb{V}_v\left(\frac{\partial \log L_i(\Theta)}{\partial \theta}\right) = \sum_{j=1}^{m} \pi_{ij}(1-\pi_{ij}) B_{ij} B_{ij}'; \tag{A6}$$

$$\mathbf{cov}_v\left(\frac{\partial \log L_i(\Theta)}{\partial q_k}, \frac{\partial \log L_i(\Theta)}{\partial q_\ell}\right) = \pi_{ik}(1-\pi_{ik}) C_{ik}^2 \mathbf{1}_{k=\ell} \quad \text{for } k,\ell = 1,2,\cdots,m; \tag{A7}$$

$$\mathbf{cov}_v\left(\frac{\partial \log L_i(\Theta)}{\partial p_k}, \frac{\partial \log L_i(\Theta)}{\partial p_\ell}\right) = \frac{\pi_{ik}(1-\pi_{ik})}{p_k^2}\mathbf{1}_{k=\ell} + \frac{\pi_{im}(1-\pi_{im})}{p_m^2} \tag{A8}$$

for $k,\ell = 1,2,\cdots,m-1$. Moreover,

$$\mathbf{cov}_v\left(\frac{\partial \log L_i(\Theta)}{\partial \theta}, \frac{\partial \log L_i(\Theta)}{\partial q_k}\right) = \pi_{ik}(1-\pi_{ik}) B_{ik} C_{ik} \quad \text{for } k = 1,2,\cdots,m; \tag{A9}$$

$$\mathbf{cov}_v\left(\frac{\partial \log L_i(\Theta)}{\partial \theta}, \frac{\partial \log L_i(\Theta)}{\partial p_k}\right) = \frac{\pi_{ik}(1-\pi_{ik})}{p_k} B_{ik} \quad \text{for } k = 1,2,\cdots,m-1; \tag{A10}$$

$$\mathbf{cov}_v\left(\frac{\partial \log L_i(\Theta)}{\partial q_k}, \frac{\partial \log L_i(\Theta)}{\partial p_\ell}\right) = \frac{\pi_{ik}(1-\pi_{ik})}{p_k} C_{ik} \mathbf{1}_{k=\ell} \tag{A11}$$

for $k = 1,2,\cdots,m$ and $\ell = 1,2,\cdots,m-1$ because

$$\mathbf{cov}_v(\lambda_{ik}, \lambda_{i\ell}) = \pi_{ik}(1-\pi_{ik})\mathbf{1}_{k=\ell} \quad \text{for } k,\ell = 1,2,\cdots,m$$

from (10) and by the multinomial property. Note that the covariance $\mathbf{cov}_v(\cdot)$ is taken over the conditional distribution of v given the observed data $\{\delta_i, \mathbb{D}_i\}_{i=1}^n$. The missing information matrix is then given by

$$\mathcal{I}_2(\Theta) = \sum_{i=1}^{n} \mathbb{V}_v\left(\frac{\partial}{\partial \Theta} \log L_i(\Theta)\right),$$

where $\mathbb{V}_v\left(\partial \log L_i(\Theta)/\partial \Theta\right)$ consists of covariance matrices (A6) to (A11) conformably with the array of parameters $\Theta = \left(\theta', q_1, \cdots, q_m, p_1, \cdots, p_{m-1}\right)'$.

References

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Boston: Kluwer Academic Publishers.

Anselin, L. and A.K. Bera (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics, in *Handbook of Applied Economics Statistics*, A. Ullah and D.E.A. Giles ed., New York: Marcel Dekker.

Case, A.C. and H.S.Rosen (1993). Budget spillovers and fiscal policy interdependence, *Journal of Public Economics*, 52, 285-307.

Cox, D.R. (1972). Regression models and life-tables, *Journal of Royal Statistical Society: Series B*, 34, 187-220.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*, New York: John Wiley & Sons, Inc.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum likelihood from incomplete data via EM algorithm, *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

Diebold, F.X., G.D. Rudebusch and D.E. Sichel (1993). Further evidence on business-cycle duration dependence, *Business Cycles, Indicators, and Forecasting*, ed. by J.H. Stock and M.W. Watson, Chicago: The University of Chicago Press.

Elbers, C. and G. Ridder (1982). True and spurious duration dependence: The identifiability of the proportional hazard model, *The review of Economic Studies*, 49:3, 403-409.

Everitt, B.S. and D.J. Hand (1981). *Finite Mixture Distributions*, New York: Chapman and Hall.

Goffe, W.L. (1996). SIMANN: A global optimization algorithm using simulated annealing, *Studies in Nonlinear Dynamics and Econometrics*, 1:3, 169-176.

Goffe, W.L., G.D. Ferrier and J. Rogers (1994). Global optimization of statistical functions with simulated annealing, *Journal of Econometrics*, 15:2, 265-287.

Guo, G. and G. Rodriguez (1992). Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala, *Journal of the American Statistical Association*, 87, 969-976.

Heckman, L. and B. Singer (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data, *Econometrica*, 52:2, 271-320.

Kelejian, H.H. and I.R. Prucha (1999). A generalized moments estimator for the autoregressive parameter in a spatial model, *International Economic Review*, 40, 509-533.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the Maximum Likelihood Estimator in the presence of infinitely many incidental parameters, *The Annals of Mathematical Statistics*, 27:4, 887-906.

Lancaster, T. (1979). Econometric methods for the duration of unemployment, *Econometrica*, 47, 939-956.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*, Cambridge: Cambridge University Press.

Lee, L.F. (2000). A numerical stable quadrature procedure, *Journal of Econometrics*, 95:1, 117-129.

Lee, L.F. (2002). Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models, *Econometric Theory*, 18, 252-277.

Lee, L.F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models, *mimeo*, Ohio State University.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society*, 44, 226-233.

Manski, C.F. (1993). Identification of Endogenous Social Effects: The Reflection Problem, *Review of Economic Studies*, 60, 531-542.

Meng, X.L. and D.B.Rubin (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm, *Journal of the American Statistical Association*, 86, 899-909.

Meyer, B.D. (1990). Unemployment insurance and unemployment spells, *Econometrica*, 58:4, 757-782.

Meyer, B.D. (1995). Semiparametric Estimation of Hazard Models, *mimeo*, Northwestern University.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm, *Journal of the Royal Statistical Society*, Series B, 61:2, 479-482.

Pinske, J. and M.E. Slade (1998). Contracting in space: An application of spatial statistics to discrete-choice models, *Journal of Econometrics*, 85, 125-154.

Pinske, J., M.E. Slade and C. Brett (2002). Spatial price competition: A semiparametric approach, *Econometrica*, 70, 1111-1153.

Prentice, R. and L. Gloeckler (1978). Regression analysis of grouped survival data with application to breast cancer data, *Biometrics*, 34, 57-67.

Sargan, J.D. (1975). Asymptotic theory and large models, *International Economic Review*, 6, 75-91.

Sueyoshi, G.T. (1995). A class of binary response models for grouped duration data, *Journal of Applied Econometrics*, 10, 411-431.

Trussell, J. and T. Richards (1985). Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure, in *Sociological Methodology 1985*, N. Tuma ed., San Francisco: Jossey-Bass, 242-276.

Van den Berg, G.J. (2000). Duration Models: Specification, Identification, and Multiple Durations, in *Handbook of Econometrics*, Vol. V, J.J. Heckman and E. Leamer ed., North-Holland: Amsterdam.