Mixture Models in Econometric Duration Analysis*

Karl Mosler
Universität zu Köln
D–50923 Köln
mosler@statistik.uni-koeln.de

Final, September 2002

Abstract

Econometric duration analysis has become an important part of methodology in econometrics, bringing forth a plenty of applications. The probability distribution of the duration of a time span is modeled through its conditional hazard rate given the covariates. When some of the covariates are unobservable, the duration, given the observable covariates, has a mixture distribution. The paper surveys and discusses recent developments in the specification, estimation, diagnosis and economic application of proportional hazard models with unobservables.

Keywords: Frailty models, mixed proportional hazard models (MPH), PH with unobservables, specification tests.

Econometric duration analysis, that is inquiring time spans between certain events and explaining them by covariates, has become an important part of econometric methodology bringing forth a plenty of applications. The probability distribution of a time span is conveniently modeled through its conditional hazard rate given the covariates. When some of the covariates are unobservable the duration, given the observable covariates, has a mixture distribution. In proportional hazard models with unobservables, called *mixed proportional hazard* (MPH) models, it is assumed that the unobserved heterogeneity (= frailty) enters the conditional hazard rate in a multiplicative way.

*Paper presented to MIXTURES 2001, Hamburg, 23.-28.7.2001. I thank two anonymous referees for their helpful suggestions concerning the models and for hints to additional literature.

There are three principal reasons to use an MPH model for analysing a given data set. As a first reason, we may suppose some generic unobserved heterogeneity in a population by which the population divides into groups which we cannot directly observe. Secondly, errors in the measured durations or covariates of a PH model may lead to an MPH model. Moreover, given a frailty model, errors in the covariates yield another frailty model. As a third reason, the baseline hazard may be of a non-simple type which can be – at least approximately – modeled as a finite mixture of simple parametric distributions. For example, if the baseline hazard decreases, its distribution may be modeled as a mixture of exponential distributions.

MPH models have been applied to problems in economics and the social sciences since some 20 years. While mixture models *per se* often are non-identified, the introduction of observable covariates into mixed duration models solves most of the identification problems; see Heckman and Taber (1994). The estimation theory of MPH models is well established and general algorithms have been developed to perform the estimation numerically. Also, many diagnostic procedures – by visual inspection or by tests – exist. A terse survey of duration analysis in general is given by Oakes (2000), while Therneau and Grambsch (2000) provide a practical introduction into various proportional hazard models and Hougaard (2000) presents methods for analysing the dependency of durations.

Recently, new tests for the classical MPH specification, which are based on stochastic order, and tests for the existence of frailty in exponential mixtures have been proposed. In estimation, MCMC and other approaches have been developed to estimate rather general shared frailty models. Concerning specification, autoregressive models have been introduced, which depict repeatedly measured durations. Applications include examples in finance, labour, and demography.

This paper surveys recent developments in the specification, estimation and diagnosis of such models and their application to problems in economics and the social sciences. In a short survey like this there is no way to cover these developments in full. Rather we will refer to the recent literature for partial surveys and concentrate on problems from an applied view.

In the sequel we consider a continuous random variable T in $[0, \infty]$ representing the duration of some time spell, a random vector X of observable covariates, and a random vector U of unobservable covariates, X and U independent. (Note that many results exist also for durations in discrete time. The mixture of exponential distributions in continuous time corresponds to the mixture of geometric distributions in discrete time.)

Section 1 introduces four mixed proportional hazard models of different generality and scope: the MPH model, the shared frailty model, a generalized form of the shared frailty model, and an autoregressive frailty model. In Section 2 several applications in economics and the social sciences are mentioned. Section 3 concerns the estimation of such models, while Section 4 is devoted to model diagnosis. Section 5 concludes.

1 Model specifications

Let $(T_i, X_i, U_i), i = 1, \ldots, n$, be an independent sample of durations T_i with their covariates. X_i is the ith row of an $n \times p$ random matrix of observed covariates and U_i is a vector of unobserved covariates, that is unobserved random effects. Assume that, conditional on $X_i = x_i$ and $U_i = u_i$, T_i has a continuous distribution with survival function $S_i(t|u_i)$, density $f_i(t|u_i)$ and hazard rate

$$h_i(t|u_i) = \frac{f_i(t|u_i)}{S_i(t|u_i)} = h_0(t)\,\gamma(u_i, x_i)\,. \tag{1.1}$$

The time dependent term $h_0(t)$ is known as the *baseline hazard rate.* The specification (1.1) is a general proportional hazard specification, saying that the covariates – observed and unobserved – enter the conditional hazard rate through a factor by which the baseline hazard is multiplied.

If γ does not depend on u_i, in particular, $\gamma(u_i, x_i) = e^{-x_i\beta}$, with some coefficient vector β, the classical *proportional hazard* (PH) arises,

$$h_i(t|u_i) = h_0(t)e^{-x_i\beta}\,, \tag{1.2}$$

which has been the working horse of duration analysis since many years; see Cox (1972, 1975). The principal advantage of this simple model consists in the decomposability of its likelihood function into two separate factors that relate to the observable covariates ("partial likelihood") and to the baseline hazard, respectively. The PH specification is rather special and can be tested in applications. Many generalizations and extensions of the model have been developed, among them PH models with unobservable heterogeneity, PH models with shared frailty, and PH models with time dependent observed covariates. With the latter, the PH model extends, in effect, to nonproportional hazard rates.

1.1 Mixed proportional hazard models

This section presents four selected hazard models which are formally related but depict different aspects of heterogeneity. Each has seen specific applications in economics.

Our first model is the mixed proportional hazard (MPH) model, which is also mentioned as the *proportional hazard* model *with unobservables* (PHU). For a comprehensive treatment, see Lancaster (1990). Choose $\gamma(u_i, x_i) = r(u_i)g(x_i)$ in (1.1) with some positive functions r and g and obtain

$$h_i(t|u_i) = h_0(t)r(u_i)g(x_i)\,. \tag{1.3}$$

As the vector u is not identifiable, but only $r(u)$, introduce the random variable $V_i = (r(U_i))^{-1}$ and write

Model I (MPH):

$$h_i(t|v_i) = h_0(t)\,\frac{g(x_i)}{v_i}\,. \tag{1.4}$$

The MPH specification implies that the survival function of T_i given the observable covariates is a mixture of survival functions, namely

$$S_i(t) = \int S_0(t)^{\frac{g(x_i)}{v}}\pi(dv),$$

where $S_0(t) = e^{-H_0(t)}$, $H_0(t) = \int_0^t h_0(s)ds$, and π is the distribution of the random variable V.

Naturally, when the distribution of V is degenerate at 1 and $g(x_i) = e^{-x_i\beta}$, the MPH reduces to the PH model, but in general the two models are different. When V^{-1} is exponential, the proportional odds model arises; see Bennett (1983) and Murphy et al. (1997).

By the MPH specification several different statistical situations can be modeled. Firstly, the unobserved heterogeneity arises from intrinsic heterogeneity in the population. Then, every individual has a specific factor v_i, which determines his or her relative hazard rate. The factor is mentioned as *frailty*, which means individual susceptibility of the event. Secondly, a multiplicative random error in the recorded durations leads to an MPH specification and, thirdly, the same holds for an additive independent random error in the observed covariates; see Lancaster (1990, pp 59ff).

In many applications of the MPH model the frailty variable V has a finite support, say $\{v_1, \ldots, v_q\}$, where q may be known or not. Then, the distribution of T_i given x_i is a *finite distribution mixture* having survival function

$$S_i(t) = \sum_{j=1}^{q} S_0(t)^{\frac{g(x_i)}{v_j}}\,\pi(v_j)\,. \tag{1.5}$$

If V takes only a finite number of values, observations divide into q groups $G^{(j)}$ according to these values, $G^{(j)} = \{i|V_i = v_j\}$ and, with $g(x_i) = e^{-x_i\beta}$, the conditional hazard rate is

$$h_i(t|v_i) = h_0(t)\ \exp(-x_i\beta - \log v_j) \quad \text{if} \ \ i \in G^{(j)}\,, \quad j = 1, \ldots, q\,. \tag{1.6}$$

We call this the *finite mixture proportional hazard* model (FMPH). Note that the groups $G^{(j)}$ are random and unknown, so the model FMPH is a *fixed effects model with unknown random groups.*

In particular, an FMPH model can be interpreted that there exist several possible "exits" to end a duration, but we do not observe which exit is realised. E.g., with two exits, the frailty variable assumes two values v_1 and v_2 and

$$S_i(t) = p\, S_0(t)^{\frac{g(x_i)}{v_1}} + (1-p)\, S_0(t)^{\frac{g(x_i)}{v_2}}\,.$$

This is a kind of competing risks model. A special case is the *latent failure-time model*, where the minimum of two independent failure times is realised; then the observed hazard can also be written as a mixture of the hazards of these failure times.

If the second exit degenerates, $v_2 \to \infty$ and $v_1 = 1$, we obtain a so-called *long-term survivor* model,

$$S_i(t) = pS_0(t)^{g(x_i)} + (1-p)\,. \tag{1.7}$$

which is the simplest competing risks model and goes back to Boag (1949). See e.g. Peng and Dear (2000) and, for a detailed treatment of long-term survivors, Maller and Zhou (1996).

1.2 Shared frailty models

Alternatively, assume that the n durations divide into q groups G_j in which the frailty is the same and that these groups are known.

More precisely, let $U = (U^{(1)}, \ldots, U^{(q)})$ be a random vector in $\mathbb{R}^q$ consisting of q iid random components and let γ in (1.1) be an exponential-linear function, such that

Model II (SF):

$$h_i(t|u^{(j)}) = h_0(t)\ \exp(-x_i\beta - u^{(j)}) \quad \text{if} \quad i \in G_j\,, \tag{1.8}$$

where $G_1, \ldots, G_q$ is a known partition of $\{1, 2, \ldots, n\}$ and β is a parameter vector to be estimated. This is the *shared frailty* (SF) model. The random variable $U^{(j)}$ is called the j-th frailty, because it influences the hazard rates in group G_j by an unknown factor which is common to all members of the group.

The shared frailty model serves to specify a certain dependency among durations: The durations in a group are dependent through the frailty variable they share. In contrast to the above finite mixture model FMPH, the SF model describes dependence within groups while the FMPH model reflects overdispersion that arises from unobserved heterogeneity of the individuals. Also, given an SF model, errors in the covariates yield a model of the same type (Li and Lin, 2000).

Formally, the SF model is a *random effects model with known groups*, the FMPH model is a *fixed effects model with random groups.* If $q = n$, we have individual frailty variables that are not shared within groups. Then the shared frailty model becomes an MPH model with not necessarily finite mixing distribution. In this sense the SF model (with $g(x_i) = \exp(-x_i\beta)$) includes the MPH model as a special case. In particular, algorithms designed to estimate the SF model can be used to estimate the MPH model as well. Further, in many applied situations one has the choice to model group effects either by random variables or by fixed coefficients. If the frailty effects in an SF model are taken as fixed effects, an ordinary PH model results.

For detailed introductions into the SF and related models and for many applications in medicine, see Hougaard (2000, esp. Chapters 7–8) and Ibrahim et al. (2001, Chapter 4). A relation of the SF model with accelerated life models used in reliability is pointed out in Bagdonavičius and Nikulin (2002, pp 43 f).

The SF model extends to the following *generalized shared frailty* model,

Model III (GSF):

$$h_i(t|u^{(j)}) = h_0(t)\ \exp(-x_i\beta - \phi(\alpha, u^{*(j)}, x_i)) \quad \text{if } i \in G_j\,, \tag{1.9}$$

and $G_1, \ldots, G_q$ is a known partition of $\{1, 2, \ldots, n\}$. Here ϕ is a given function, $U^{*(1)}, \ldots U^{*(q)}$ denote iid random vectors in $\mathbb{R}^k$ and α is a parameter vector with $var(\phi(\alpha_0, u^{*(j)}, x_i)) = 0$ at some α_0. The SF model assumes (1.9) with $U^{(j)} = \phi(\alpha, U^{*(j)}, x_i)$.

In the generalized SF model the frailty of an individual depends on its group frailty $U^{*(j)}$ and, in addition, on its individual covariates; this functional dependence is parameterized by α. The model is rather flexible and lends itself to a stepwise estimation of α and β, given temporary values of the frailties $U^{*1}, \ldots, U^{*q}$; see Section 3 below.

1.3 An autoregressive frailty model

Most economic and social science data are observational data, many of them arising from time series. Then, successively observed durations are often dependent over time. As a fourth model we introduce an *autoregressive frailty* (ARF) model for repeated observations. It depicts unobserved heterogeneity among individuals having recurrent events. For each individual a series of time spells is observed over time. Let T_{ik} denote the k-th duration of individual i, $k = 2, 3, \ldots$, and assume that its hazard rate, conditional on the past durations $T_{i,k-1}, \ldots, T_{i1}$ and a frailty variable V_i, is

Model IV (ARF):

$$h_{ik}(t|t_{i,k-1}, \ldots, t_{i1}, v_i) = h_0(t) \, \frac{1}{v_i \, \psi_{ik}} \, ,$$

where

$$\psi_{ik} = \psi(\mu_{ik}) \, , \qquad \mu_{ik} = E(T_{ik}|t_{i,k-1}, \ldots, t_{i1}) \, .$$

That is, to obtain the conditional hazard rate at some time, given the past observations, the baseline hazard rate is divided by the frailty, which is an individual effect, and an additional factor that depends on the conditional mean duration. Then, conditional on the past observations, the duration T_{ik} has a mixture distribution with survival function

$$S_{ik}(t|t_{i,k-1}, \ldots, t_{i1}) = \int S_0(t)^{\frac{1}{v \, \psi_{ik}}} \, \pi(dv|t_{i,k-1}, \ldots, t_{i1}) \, .$$

Obviously, when ψ is constant, the ARF reduces to the MPH model without covariates.

Especially with a Weibull baseline hazard $h_0(t) = a \, t^{a-1}$ we get $S_0(t) = \exp(-t^a)$ and the conditional Weibull mixture

$$S_{ik}(t|t_{i,k-1}, \ldots, t_{i1}) = \int \exp\left(-\frac{t^a}{v \, \psi_{ik}}\right) \, \pi(dv|t_{i,k-1}, \ldots, t_{i1}) \, .$$

Further,

$$E(T_{ik}|t_{i,k-1}, \ldots, t_{i1}, v_i) = \Gamma\left(1 + \frac{1}{a}\right) \cdot (v_i \, \psi_{ik})^{1/a} \, .$$

We obtain that, given the past durations,

$$T_{ik} = \Gamma\left(1 + \frac{1}{a}\right) \cdot \psi_{ik}^{1/a} \cdot \epsilon_{ik} \tag{1.10}$$

holds, where ϵ_{ik} follows a mixed Weibull distribution with parameters a and $\frac{1}{V}$, V being distributed with $\pi(\cdot|t_{i,k-1}, \ldots, t_{i1})$. An important special case is $a = 1$ and $\psi_{ik} = \mu_{ik}$. Then, T_{ik} is an exponential mixture scaled by the conditional mean duration.

So far, we have not yet fixed the dependence of μ_{ik} on past durations. For this, an $ARMA(p,q)$ specification presents itself:

$$\mu_{ik} = \alpha_0 + \sum_{j=1}^{p} \alpha_j x_{i,k-j} + \sum_{j=1}^{q} \beta_j \mu_{i,k-j} \,,$$

with some $\alpha_j, \beta_j \in \mathbb{R}$ such that

$$E(T_{ik}) = \frac{\alpha_0}{1 - \sum_{j=1}^{p} \alpha_j + \sum_{j=1}^{q} \beta_j} \,.$$

Then, the durations of individual i are described by an $ARMA(p,q)$ process

$$T_{ik} = \alpha_0 + \sum_{j=1}^{p \vee q} (\alpha_j + \beta_j) x_{i,k-j} - \sum_{j=1}^{q} \beta_j \eta_{i,k-j} + \eta_{ik} \,,$$

where $\eta_{ik} = T_{ik} - \mu_{ik}$ is a non-Gaussian innovation sequence and $p \vee q = \max\{p,q\}$.

2 Applications in Economics and Social Sciences

In this section we shall mention several examples in which the MPH and shared frailty models have been applied to problems in economics and the social sciences.

1. Davies (1993) investigated data on residential mobility in a given area. Residential mobility of a certain household type is observed to decline with time. He showed that this occurs partly due to an unobserved frailty effect (latent propensity to move) by which high-frequency movers become progressively underrepresented in the population.

2. A classical application of MPH models on economic data concerns the duration of individual unemployment. Flinn and Heckman (1982) and many subsequent authors have investigated the length of unemployment spells with socio-economic covariates of the job seeking individual. These authors argued (and were able to support by data) that besides observed covariates like age, education, health and experience further variates, which are unobserved, determine the duration of unemployment.

3. Another popular application field is the analysis of fecundability in demography. Several authors have investigated durations in the process of human reproduction.

Recently, Li and Choe (1997) have analyzed the probability and the timing of second births in China. Observed covariates are the mother's age, education and living characteristics, the sex of her first child and whether she has accepted a so-called one-child certificate. The authors assume a two component mixture duration (1.7), where one component corresponds to the occurrence and the other component to the non-occurrence of second births,

$$S_i(t) = p\, S_0^{g(x_i)}(t) + (1-p)\,.$$

p is the probability of a second birth, and $S_1(\cdot|x)$ is the conditional survival function of the duration to second birth, if there is one. p is estimated by a logistic regression and S_1 by a Cox regression with piecewise linear baseline hazard.

4. Earlier applications of the MPH model in demography are Heckman and Walker (1990) who analyzed the waiting times until first conception by an MPH model and Behrman et al. (1992) who studied the economic determinants of mortality. See also Sickles and Taubman (1986) for individual heterogeneity in a model of leisure and morbidity.

   Shared frailty models have been employed in analysing the mortality of twins (Yashin and Iachine (1995)).

5. Lindeboom and Kerkhofs (2000) have employed proportional hazard models with unobserved heterogeneity to explain sickness absenteeism of school teachers. Durations are sickness and work spells of teachers at Dutch primary schools. Observed covariates include the teacher's age, gender and contract type and the school's ability to replace absent workers, among others. As large differences in average spell length are observed between schools, a shared frailty model is assumed, where each school forms a group. The data consist of more than 21 000 spells of some 5 000 teachers at 426 schools. These authors do not estimate the shared MPH model with random frailties but employ three models in which the frailty of each school is estimated as a fixed effect.

6. Further economic applications of MPH models include the duration of strikes in the US manufactoring industry (Kennan, 1985; Jaggia, 1991). Here, observed covariates are business cycle effects on industrial production, time squared and monthly dummy variables. See also the study of Brännäs and Rosenqvist (1988) on the incidence of coffee purchases by consumers.

7. A recent field of application is analysing the microstructure of financial markets. It is a common practice to model the distribution of stock prices – slightly asymmetric with fat tails – by finite mixtures of normal distributions. Also

the durations between certain market events, such as the time it takes until the prices of stocks change by a given amount, may be conveniently modeled by finite mixtures of exponential distributions.

Engle and Russell (1998) (see also Engle (2000)) used the ARF model with constant frailty to investigate certain durations observed with a stock: one is the time spell between trades, and the other is the duration of a price change.

3 Estimation

Recent advances have been made in the estimation of all four models. They concern (1) the introduction of fully non-parametric estimators, (2) the use of MCMC procedures which appear to be robust against misspecifications of the model and (3) the availability of software for the analysis of proportional hazard models and of finite mixtures.

Model I, the classical MPH model,

$$h_i(t|v_i) = h_0(t)\ \exp(-x_i\beta - \log v_i)\,, \tag{3.1}$$

is the oldest. Besides the parameter vector β, the distribution function F_V of V and the baseline hazard h_0 are unknown. Many authors have contributed to their estimation, starting from Lancaster (1979). They differ in that they impose parametric assumptions either on F_V or h_0 or both. Most often F_V is assumed to be Gamma or normal, and h_0 to be Weibull or exponential. While Lancaster (1979) assumes that both belong to known parametric families, Heckman and Singer (1984b) estimate F_V nonparametrically and h_0 parametrically; see also Meyer (1990) for discrete approximation. However, in applications of the PH model (Cox, 1972; Heckman and Singer, 1984a) the baseline hazard has often proved to have an irregular shape, which asks for a nonparametric estimation. Such estimators, combined with parametric estimators of F_V, have been given by Murphy (1995) and Nielsen et al. (1992); see also Horowitz (1998, pp 156 ff).

It has been early stated ((Elbers and Ridder, 1982) that the MPH model is nonparametrically identified if the frailty variable has finite first moment and the observables are not all constant. Horowitz (1999) developed a fully nonparametric estimator for both the baseline hazard h_0 and the distribution of V. He noted that the MPH model can be equivalently formulated as the following transformation model:

$$\log H_0(T_i) = X_i\beta + \log V + \epsilon\,, \tag{3.2}$$

where ϵ is a random variable having an extreme value distribution, $P[\epsilon > t] = \exp(e^{-t})$, independent of V, and $H_0(t) = \int_0^t h_0(s)ds$ denotes the cumulative baseline hazard. Using kernel estimators for $\frac{1}{|\beta_1|} \log H_0(\cdot)$ and for the distribution of $\frac{1}{|\beta_1|}(\log V + \epsilon)$, Horowitz (1999) solves the nonparametric estimation problem.

Bayesian sampling methods like data augmentation – a special case of the Gibbs sampler – and MCMC provide a useful and widely applied methodology to estimate frailty and MPH models; see Diebolt and Robert (1994) and the survey by Sinha and Dey (1997). These methods appear to be versatile and robust. However, the algorithms are computer intensive and have to be implemented with care.

E.g., for the general shared frailty model (Model III) Jalaluddin and Kosorok (2000) use the Metropolis-Hastings algorithm in connection with the bootstrap. In each step the parameters α and β are updated, given the iid random vectors u_j, and the working partial joint likelihood of α, β and $u_1, \dots, u_q$ is calculated. Their algorithm consists of two major steps: a main Markov chain simulation based on the entire data to estimate α and β (by the mode vector) and a series of Markov chain simulations based on bootstrap samples to estimate the covariance matrix from the simulated mode vectors. To calculate acceptance probabilities, Jalaluddin and Kosorok (2000) use an efficient data structure in which the contributions of the different working parameters α, β and the u_j to the probability are separately updated. The validity of the procedure rests mainly on the iid property of the u_j. Thus, the procedure appears to be robust against misspecifications of h_0 and ϕ. Naturally, since the approach is computer intensive, the number of groups q or – in the MPH case – of individuals n is limited. Jalaluddin and Kosorok (2000) present a numerical example with $n = 38$.

When parts of the data are censored, a point process formulation (Aalen, 1978; Nielsen et al., 1992; Andersen et al., 1993) of the model is useful, Kuo and Peng (1995) have estimated the FMPH model under censoring with different EM and MCMC approaches. For a fully nonparametric estimation of the MPH model with right censored data, see Gørgens and Horowitz (1999). Lemdani and Pons (1997), by martingale methods, derived asymptotic properties of the ML estimate for the finite MPH model with censored data.

To estimate the ARF model, the equivalent formulation (1.10) is considered. It can be estimated by ARCH methods and partially employing standard ARCH computer code; see Engle and Russell (1998) and Engle (2000) for the case when $var(V) = 0$.

During the past years the availability of standard software for duration and mixture analysis has greatly improved. (By a standard software package we mean a computer code – either free or commercial – which is used by a large number of users and has virtually no errors.)

All large commercial software packages (like S-Plus, SAS, SPSS and STATA) contain computer codes for the analysis of the classical proportional hazard model, some of them also for the shared frailty model. They can be used as moduls for the analysis of similar MPH specifications.

Several specialized software packages have been developed for the analysis of finite mixtures. Recent information on most of them is found in the appendix of McLachlan and Peel (2000). Haughton (1997) has reviewed and compared five such packages. The monograph by Böhning (2000) includes a detailed introduction into the author's computer code C.A.MAN.

The estimation of the MPH model is rather non-robust with respect to parametric assumptions on the frailty distribution π (Heckman and Singer, 1982). But also the nonparametric estimation of π by the EM algorithm appears to be very sensitive to the choice of initial values. While the theory of parametric and nonparametric estimation of the MPH model (also with censoring) has been largely established, the implementation of these methods can pose severe problems.

The EM algorithm is a natural, well investigated and widely applied device to estimate models with unobserved covariates; see McLachlan and Krishnan (1997) and McLachlan and Peel (2000). As the EM algorithm is rather slow, several variants have been proposed to increase its speed; see Aitkin and Aitkin (1996). But its convergence, particularly in duration analysis, may heavily depend not only on the chosen starting values but also on the stopping rule. See Seidel et al. (2000a) who investigated its behavior in estimating a finite mixture of exponential distributions. EM variants employing minimum message length (MML) appear to be more robust with respect to starting values (Hansen and Yu, 2001).

As an alternative to the EM algorithm, several authors (see Sickles and Taubman (1986); Huh et al. (1998)) have proposed to estimate the MPH and, more general, the shared frailty model within the framework of a penalized Cox model. In particular, Therneau and Grambsch (2000, Chapter 9) assume either a gamma frailty distribution or a Gaussian random effects model; in estimating this, standard S-Plus code can be employed. Another feasible alternative is the nonparametric ML estimation of the mixing distribution (Seidel and Ševčíková, 2002).

In general, methods are needed together with standard implementations which are portable and allow results which can be reproduced by other researchers. In particular, with computer intensive methods of estimation like the MCMC, *the implementation is the estimator* so that any statements about properties of the estimator are statements about its specific implementation. Some remedy is seen to come from the increasing use of standard software packages, commercial or otherwise publicly distributed.

4 Diagnosis

Many diagnostic procedures have been proposed in the literature to check the validity of PH model specifications. These are visual and graphical procedures on one hand, and specification tests on the other hand. The latter procedures test for the existence of unobserved heterogeneity and – if a finite MPH model is assumed – the number of mixture components.

Some graphical checks of proportional hazard assumptions are collected in Chapter 11 of Klein and Moeschberger (1997); see also Cox (1979) and Kay (1984).

The simplest case is the MPH model without covariates and with a known baseline hazard. It can – via the baseline distribution function – be transformed into a mixture of exponentials model. To test the null hypothesis of a non-mixed exponential distribution against a mixture of exponential distributions, a number of tests is available; for a comparison see Mosler and Seidel (2001).

In the sequel we survey three principal test approaches to test for MPH or frailty specifications. Then we present a procedure to test for necessary conditions which are derived from the MPH model. The search for alternative specifications, when the test rejects MPH, is illustrated by a possible nonproportional hazard specification, the event contingent hazard model.

4.1 Score, likelihood ratio and goodness-of-fit tests

In general, tests for the existence of mixture heterogeneity can be divided into three groups: (1) score tests for overdispersion, (2) likelihood ratio tests, and (3) goodness-of-fit tests.

(1) Testing for homogeneity in mixture models, that is for $var(V) = 0$, is closely related to the problem of detecting overdispersion in a one-parameter exponential family of distributions. The variance of such a distribution is determined by its mean. Further, any mixture of distributions from a one-parameter exponential family is a dilation of the non-mixed distribution which has the same mean as the mixture (Shaked (1980)). Therefore a mixed distribution has always larger variance than the non-mixed distribution with the same mean.

Score tests that detect overdispersion have been developed in parametric and nonparametric proportional hazard settings: Jaggia and Trivedi (1994) and Jaggia (1997) propose score tests for MPH under right censoring but without covariates; they assume $h_0(t) = const$, that is, a mixture of exponentials model. Commenges and Andersen (1995) (see also Klein and Moeschberger (1997, Chapter 13)) present a nonparametric test for the shared frailty model with covariates which is based on the partial likelihood.

(2) Likelihood ratio (LR) tests are used for testing the null hypothesis of homogeneity against a finite mixture (with known number $q \geq 2$) of components. LR tests are suited also to test for the null of $p \geq 2$ components against an alternative of $q \geq p+1$ components. In practice, these tests are used repeatedly to determine the number of mixture components. It is well known that the usual chi-square asymptotics do not hold. The LR statistic has a rather complicated distribution. Chen et al. (2001) recommend a modified LR statistic that has a simple limiting distribution and is – in certain situations – convenient to apply.

In most cases, however, the critical quantiles must be determined numerically for a given distribution. In exponential mixtures, as shown by Seidel et al. (2000a), the numerics of such tests can be very problematic: The test statistic and the critical quantile depend heavily on the chosen implementation of the EM algorithm. Seidel et al. (2000b) calculated the power of LR tests for mixtures of exponentials under different implementations and demonstrate that "global" – that is multistart – maximization of the likelihood function does not result in the largest power.

(3) The third test approach is goodness-of-fit. To test for the MPH model without covariates and with a known baseline hazard – in fact, for a mixture of exponentials model – Mosler and Seidel (2001) investigate several goodness-of-fit tests and compare their power with a score test for overdispersion and an LR test. It comes out that a proper combination of the score test with a goodness-of-fit test (the Anderson-Darling test) has the best overall power. Two new goodness-of-fit tests for this problem are given in Baringhaus and Henze (2000).

4.2 Testing for consequences of the MPH model

Now, assume that the MPH model (1.4) holds and that $E(V) = 1$. A central question is whether, under the MPH assumption, V has a positive variance. Otherwise V is a constant and equal to 1, and there exists no unobserved heterogeneity.

Firstly, consider the classical Cox test. If $var(V) = 0$ then for any given x_i and x_j the hazard ratios are constant over time,

$$\frac{h_i(t)}{h_j(t)} = \frac{g(x_i)}{g(x_j)}, \tag{4.1}$$

hence,

$$\frac{d}{dt}\left(\frac{h_i(t)}{h_j(t)}\right) = 0 \quad \text{for all } t\,. \tag{4.2}$$

Cox (1972) proposed a simple test for (4.2) by using artificial time-dependent covariates. While the Cox test is useful to decide whether the PH specification is correct, it cannot be used for testing the more general MPH specification. Namely (Mosler and Philipson, 1999), under the MPH model,

$$\text{(4.2) implies that either } var(V) = 0 \text{ or } g(x_i) = g(x_j)\,.$$

If, under the MPH assumption, the Cox test rejects (4.2), this gives no evidence against the MPH model. Rather, it suggests the presence of unobserved heterogeneity.

Secondly, the MPH specification implies (Mosler and Philipson (1999)) that any two individual survival functions are ordered in one or the other direction,

$$S_i(t) \geq S_j(t) \quad \text{for all } t \quad \text{or} \quad S_i(t) \leq S_j(t) \quad \text{for all } \ t\,. \tag{4.3}$$

Note that this order restriction (4.3) is weaker than – that is, implied by – the above Cox restriction (4.2). The order restriction (4.3) can be visually checked and also tested as a null hypothesis by two recently developed tests for stochastic order in either direction; see Schmid and Trede (1996), Rothan and Gideon (1996), and Schmid and Kraft (2002). One test statistic is of Kolmogorov-Smirnov type (minimum of suprema of empirical distribution functions) while the other is an area statistic (minimum of certain areas in the *p*-*p*-plot).

A further necessary restriction is that, under MPH, any two individual durations must have the same support (Mosler and Philipson (1999)). The equality of supports may be tested as a null hypothesis, too.

The question remains what is to be done when either the null hypothesis of ordered survivals or of equal support is rejected. Many alternatives to the MPH model may be hypothesized. One of them, which occurs frequently with observational data, is that there exist historical events which affect the observed durations. As the individual durations start at different calendar times, a historical event will influence them in a different way. In the following *event contingent hazard* ECH model, an event at time τ_0 is introduced that influences the observed durations after it has occurred. The hazard rate conditional on x_i and the additional covariate τ_0 is assumed to be

$$h_i(t|\tau_i) = h_0(t)g(x_i)[1 + c(t + \tau_i - \tau_0)]\,,$$

where

$c:$ $\mathbb{R} \to \mathbb{R}$ with $c(s) = 0$ if $s \leq 0$,

$\tau_0:$ calendar time of a historical event,

$\tau_i:$ calendar starting time of the i-th duration (= additional observed covariate).

If a duration T_i starts at calendar time τ_i, the event becomes relevant to it when $t \geq \tau_0 - \tau_i$. The term $1 + c(t + \tau - \tau_0)$ represents the multiplicative effect of the event on the hazard rate, while h_0 is an event independent baseline hazard.

It can be shown that the ECH model is identified and that the tests for ordered survivals are consistent against the ECH model.

5 Conclusions

To summarize we list the main statements of this survey:

- The MPH and the shared frailty models have important applications in economics and social sciences:

  Labour Economics: durations of education, job and unemployment

  Demography: times until birth, death

  Financial Markets: spell between transactions, price movements

  Life insurance and pension funds: evolution of a contract in time

- To estimate the models, also with censored data, well investigated general procedures are available. But not many implementations, especially with censoring, nor comparisons of methods have been published.

- While there exist various procedures for diagnosis, there is a lack of alternative models when a diagnostic test rejects the MPH model.

- Likelihood estimation and testing in mixtures of exponential and other life distributions give rise to severe numerical problems.

- Portability of an implementation is an important issue. In principle, without it no results can be reproduced by other researchers. Existing standard software should be used whenever possible.

Finally, the question remains whether and when it is “worth the risk” to employ such models in place of simpler proportional hazard specifications. Particularly, what are the criteria for a successful application of the finite MPH model? If a generic population heterogeneity is present, an application may be called successful if after estimation of the model the groups can be interpreted in a sensible way and, at least, a partial classification of individuals into groups is possible. If not, data fit should be the principal criterion.

References

AALEN, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics* **6**, 701–726.

AITKIN, M. and AITKIN, I. (1996). A hybrid EM/Gauss — Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing* **6**, 127–130.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes.* Springer Verlag, New York.

BAGDONAVIČIUS, V. and NIKULIN, M. (2002). *Accelerated Life Models.* Chapman & Hall, London.

BARINGHAUS, L. and HENZE, N. (2000). Tests of fit for exponentiality based on a characterization via the mean residual life function. *Statistical Papers* **41**, 225–236.

BEHRMAN, J., SICKLES, R. C., TAUBMAN, P. and YAZBECK, A. (1992). Black-white differentials in mortality: Results from the retirement history survey. *Journal of Econometrics* **50**, 183–204.

BENNETT, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273–277.

BOAG, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B* **11**, 15–53.

BÖHNING, D. (2000). *Computer-Assisted Analysis of Mixtures and Applications.* Chapman & Hall, Boca Raton, Florida.

BRÄNNÄS, K. and ROSENQVIST, G. (1988). Semiparametric estimation of heterogenous count data models. Working Papers 187, Swedish School of Economics and Business Adminstration.

CHEN, H., CHEN, J. and KALBFLEISCH, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society, Series B* **63**, 19–29.

COMMENGES, D. and ANDERSEN, P. K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis* **1**, 145–156.

COX, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

COX, D. R. (1979). A note on the graphical analysis of survival data. *Biometrika* **66**, 188–190.

DAVIES, R. B. (1993). Nonparametric control for residual heterogeneity in modelling recurrent behaviour. *Computational Statistics & Data Analysis* **16**, 143–160.

DIEBOLT, J. and ROBERT, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society, Series B* **56**, 363–375.

ELBERS, C. and RIDDER, G. (1982). True and spurious duration dependence: The identifiability of the proportional hazard model. *Review of Economic Studies* **49**, 403–410.

ENGLE, R. F. (2000). The econometrics of ultra-high-frequency data. *Econometrica* **68**, 1–22.

ENGLE, R. F. and RUSSELL, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* **66**, 1127–1162.

FLINN, C. and HECKMAN, J. (1982). New methods for analyzing structural models of labor force dynamics. *Journal of Econometrics* **18**, 115–168.

GØRGENS, T. and HOROWITZ, J. L. (1999). Semiparametric estimation of a censored regression model with an unknown transformation of the dependent variable. *Journal of Econometrics* **90**, 155–191.

HANSEN, M. H. and YU, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association* **96**, 746–774.

HAUGHTON, D. (1997). Packages for estimating finite mixtures: A review. *The American Statistician* **51**, 194–205.

HECKMAN, J. J. and SINGER, B. (1982). The identification problem in econometric models for duration data. In W. Hildenbrand, ed., *Advances in Econometrics*, 39–77. Cambridge University Press, Cambridge.

HECKMAN, J. J. and SINGER, B. (1984a). The identifiability of the proportional hazard model. *Review of Economic Studies* **51**, 231–241.

HECKMAN, J. J. and SINGER, B. (1984b). A method for minimising the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52**, 271–320.

HECKMAN, J. J. and TABER, C. R. (1994). Econometric mixture models and more general models for unobservables in duration analysis. *Statistical Methods in Medical Research* **3**, 279–299.

HECKMAN, J. J. and WALKER, J. R. (1990). Estimation fecundability from data on waiting times to first conception. *Journal of the American Statistical Association* **85**, 283–294.

HOROWITZ, J. L. (1998). *Semiparametric Methods in Econometrics.* Springer, New York.

HOROWITZ, J. L. (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica* **67**, 1001–1028.

HOUGAARD, P. (2000). *Analysis of Multivariate Survival Data.* Springer, New York.

HUH, K., POSTER, A. K. and SICKLES, R. C. (1998). Maximum penalized likelihood estimation of mixed proportional hazard models. *Communications in Statistics - Theory and Methods* **27**, 2143–2164.

IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2001). *Bayesian Survival Abnalysis.* Springer, New York.

JAGGIA, S. (1991). Specification tests based on the heterogeneous generalized gamma model of duration: With an application to Kennan's strike data. *Journal of the Applied Econometrics* **6**, 169–180.

JAGGIA, S. (1997). Alternative forms of the score test for heterogeneity in a censored exponential model. *Review of Economics and Statistics* **79**, 340–343.

JAGGIA, S. and TRIVEDI, P. K. (1994). Joint and seperate score tests for state dependence and unobserved heterogeneity. *Journal of Econometrics* **60**, 273–291.

JALALUDDIN, M. and KOSOROK, M. R. (2000). An algorithm for robust inference for the Cox model with frailties. *Journal of Computational and Graphical Statistics* **9**, 642–652.

KAY, R. (1984). Goodness-of-fit methods for the proportional hazards model: A Review. *Reviews of Epidemiology Santè Publications* **32**, 185–198.

KENNAN, J. (1985). The duration of contract strikes in U.S. manufacturing. *Journal of Econometrics* **28**, 5–28.

KLEIN, J. and MOESCHBERGER, M. (1997). *Survival Analysis: Techniques for Censored and Truncated Data.* Springer, New York.

KUO, L. and PENG, F. (1995). A mixture-model approach to the analysis of survival data. Unpublished mimeo.

LANCASTER, T. (1979). Econometric methods for the duration unemployment. *Econometrica* **47**, 939–956.

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

LEMDANI, M. and PONS, O. (1997). Estimation and tests in finite mixture models for censored survival data. *Statistics* **29**, 363–388.

LI, L. and CHOE, M. K. (1997). A mixture model for duration data: Analysis of second births in China. *Demography* **34**, 189–197.

LI, Y. and LIN, X. (2000). Covariate measurement errors in frailty models for clustered survival data. *Biometrika* **87**, 849–866.

LINDEBOOM, M. and KERKHOFS, M. (2000). Multistate models for clustered duration data — An application to workplace effects on individual sickness absenteeism. *Review of Economics and Statistics* **82**, 668–684.

MALLER, R. A. and ZHOU, X. (1996). *Survival Analysis with Long-Time Survivors.* J. Wiley, Chichester.

MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions.* J. Wiley, New York.

MCLACHLAN, G. J. and PEEL, D. (2000). *Finite Mixture Models.* Wiley & Sons, New York.

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

Mosler, K. and Philipson, T. (1999). Specification analysis in mixed hazard models and a test of crossing survival functions. *Statistical Papers* **40**, 37–54.

Mosler, K. and Seidel, W. (2001). Testing for homogeneity in an exponential mixture model. *Australian and New Zealand Journal of Statistics* **43**, 231–247.

Murphy, S., Rossini, A. and van der Vaart, A. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* **92**, 968–976.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *Annals of Statistics* **23**, 182–198.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19**, 25–43.

Oakes, D. (2000). Survival analysis. *Journal of the American Statistical Association* **95**, 282–285.

Peng, Y. and Dear, K. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56**, 237–243.

Rothan, S. A. M. and Gideon, R. A. (1996). A two-sample distribution-free scale test of the Smirnov type. *Communications in Statistics - Simulation and Computation* **25**, 785–800.

Schmid, F. and Kraft, S. (2002). Nonparametric tests based on area statistics. *Communications in Statistics - Simulation and Computation* .

Schmid, F. and Trede, M. (1996). Testing for first order stochastic dominance in either direction. *Computational Statistics* **11**, 165–173.

Seidel, W. and Ševčíková, H. (2002). Efficient calculation of the NPMLE of a mixing distribution for mixtures of exponentials. *UniBw Hamburg, Discussion Papers in Statistics and Quantitative Economics* **96**.

Seidel, W., Mosler, K. and Alker, M. (2000a). A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics* **52**, 481–487.

Seidel, W., Mosler, K. and Alker, M. (2000b). Likelihood ratio tests based on subglobal optimization: A power comparison in exponential mixture models. *Statistical Papers* **41**, 85–98.

Shaked, M. (1980). On mixtures from exponential families. *Journal of the Royal Statistical Society, Series B* **42**, 192–198.

Sickles, R. C. and Taubman, P. (1986). An analysis of the health and retirement status of the elderly. *Econometrica* **54**, 1339–1356.

Sinha, D. and Dey, D. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* **92**, 1195–1212.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data.* Springer, New York.

Yashin, A. and Iachine, I. (1995). Genetic analysis of durations: Correlated frailty model applied to survival of Danish twins. *Genetic Epidemiology* **12**, 529–538.