Multivariate Survival Models

Germán Rodríguez
grodri@princeton.edu

Spring, 2001; revised Spring 2005

In this unit we study models for *multivariate* survival (in the statistical sense of many outcomes, not just many predictors).

1 Areas of Application

We start by reviewing four main areas of applications of these models.

1.1 Series of Events

One area of interest is processes where each individual may experience a succession of events. Examples include birth intervals and spells of unemployment.

Because the various events occur to the same individual, the waiting times will in general not be independent. Some couples tend to have short birth intervals while others have long ones. Observed covariates such as contraceptive use may explain some of the association. In general, however, there will remain some correlation due to unobserved individual traits.

Because the events occur one after the other, it will generally be the case that only the last interval can be censored. This introduces some simplification in estimation. In particular, it makes it possible to study the sequence using *successive conditioning*.

To fix ideas consider an example with three intervals. The joint density of T_1, T_2 and T_3

$$f_{123}(t_1, t_2, t_3)$$

can always be written as the product of the marginal of T_1, the conditional distribution of T_2 given T_1, and the conditional distribution of T_3 given T_1 and T_2:

$$f_1(t_1)\, f_{2|1}(t_2|t_1)\, f_{3|12}(t_3|t_1, t_2).$$

These two equations are an identity: any joint distribution can be factored in this way.

A typical contribution to the likelihood function, given the fact that T_3 is the only waiting time that can be censored, will look like

$$f_1(t_1)\, f_{2|1}(t_2|t_1)\, \lambda_{3|12}(t_3|t_1,t_2)^{d_3} S_{3|12}(t_3|t_1,t_2).$$

where the last term is, as usual, the conditional survival function for censored cases and the conditional density for deaths.

As long as we model the conditional distributions using different parameters, the likelihood will factor into separate components. Typically, one would model the conditional distributions by introducing the previous waiting times as covariates, for example we could write the three-equation model

$$\begin{aligned}
\lambda_1(t_1|x) &= \lambda_{01}(t_1)e^{x'\beta_1} \\
\lambda_2(t_2|t_1,x) &= \lambda_{02}(t_2)e^{x'\beta_1+s_1(t_1)} \\
\lambda_3(t_3|t_1,t_2,x) &= \lambda_{03}(t_2)e^{x'\beta_1+s_1(t_1)+s_2(t_2)}
\end{aligned}$$

where $s(t)$ denotes a smooth term on t, such as a smoothing or regression spline.

My own work on birth intervals (Rodríguez et al. 1984) used this approach. It turns out that only the previous interval turned out to be relevant, so our models were simplified. Some of the advantages of this approach are

- It is easy, because it breaks down into a series of univariate analyzes.
- It is consistent with sequential decision making, where the actual values of $t_1,\ldots,t_{j-1}$ may affect behavior influencing t_j.

On the other hand it has the disadvantage of using separate parameters for each spell.

1.2 Kindred Lifetimes

A second area where we may use multivariate survival models consists of related lifetimes, such as the survival of husband and wife, siblings, or other kin. Following Vaupel I will call these *kindred* lifetimes. In general there is reason to believe that these lifetimes are correlated, because of common unobserved characteristics of the couple (in the case of husband and wife survival) or the family (in the case of sibling survival).

An important feature of kindred lifetimes is that any (or all) of the waiting times may be censored. With three children, for example, you may observe 8 different patterns of censoring. This means that we cannot adopt the simple sequential approach outlined earlier for series of events, as we will often lack the information needed. For example we can't very well model T_2 as a function of T_1 when T_1 is censored, at least not in the simple way we have described. Thus, we need a more general approach.

1.3 Competing Risks

We have already encountered a third type of multivariate data in our discussion of competing risks, where $T_1, T_2, \ldots, T_k$ represent *latent* survival times to different causes of death.

As noted earlier, estimation of these models is complicated by the fact that we only observe

$$T = \min\{T_1, \ldots, T_k\}$$

and even this can be censored. Keep in mind, however, that the models that follow could be used, at least conceptually, in this context.

1.4 Event History Models

The fourth and final type of multivariate data involves transitions among several types of states. This combines elements of competing risk models with models for series of events.

Consider for example the analysis of nuptiality. You start in the single state. From there you can move to cohabiting or married. From cohabiting you can move to married or to separated. And so on. If you distinguish separations from marriage or cohabiting as well as widowhood and divorce, you probably have about 15 possible transitions of interest.

Analysts often study one type of transition, for example age at first marriage or marriage dissolution. With event history data, however, one may study the complete process, allowing for inter-dependencies among the different kinds of transitions. The nature of the data allows conditioning each move on the entire history of previous moves.

A closely related subject in demography are multi-state models. A lot of work in that area assumes a homogeneous population with constant transition rates and independent moves, and emphasizes analytic results, such as the steady-state proportion in each state. In some ways event history models are to multi-state models what Cox regression models are to the traditional life table.

2 Bivariate Survival Models

We consider first the case of only two survival times, T_1 and T_2. This section follows Cox and Oakes (1984, Chapter 10) and Guo and Rodríguez (1992).

2.1 Basic Definitions

Interest will focus on the *joint* survival

$$S_{12}(t_1, t_2) = \Pr\{T_1 \geq t_1, T_2 \geq t_2\}.$$

Note that $S_{12}(t, t)$ is the probability that both units are alive at t.

We also have the *marginal* survival function

$$S_1(t_1) = \Pr\{T_1 \geq t_1\} = S_{12}(t_1, 0),$$

and similarly for t_2. If T_1 and T_2 were independent the joint survival function would be the product of the marginals.

We might also be interested in the *conditional* survival function, which has two variants

$$S_{1|2}(t_1|T_2 = t_2) = \Pr\{T_1 \geq t_1|T_2 = t_2\},$$

giving the survival probabilities given that the other unit failed at time t_2, and

$$S_{1|2}(t_1|T_2 \geq t_2) = \Pr\{T_1 \geq t_1|T_2 \geq t_2\},$$

given that the other unit survived to just before time t_2.

Associated with each of these survival functions there will be a cumulative hazard function, which can be obtained by taking minus the log of the survival function. There will also be a hazard function, which can be obtained by taking derivatives of the cumulative hazard.

Specifically, we can define the *joint* hazard function as

$$\lambda_{12}(t_1, t_2) = \lim_{dt \to 0} \Pr\{T_1 \in [t_1, t_1 + dt), T_2 \in [t_2, t_2 + dt)|T_1 \geq t_1, T_2 \geq t_2\}/dt^2.$$

This is the instantaneous risk that one unit fails at t_1 and the other fails at t_2 given that they were alive just before t_1 and t_2, respectively.

We can also define a *marginal* hazard,

$$\lambda_1(t_1) = \lim_{dt \to 0} \Pr\{T_1 \in [t_1, t_1 + dt)|T_1 \geq t_1\}/dt.$$

Under independence, the joint hazard is the sum of the marginal hazards. Can you prove this result?

We can also define *conditional* hazards

$$\lambda_{1|2}(t_1|T_2 = t_2) = \lim_{dt \to 0} \Pr\{T_1 \in [t_1, t_1 + dt) | T_1 \geq t_1, T_2 = t_2\}/dt$$

which tracks the risk for one unit given that the other failed at t_2, and

$$\lambda_{1|2}(t_1|T_2 \geq t_2) = \lim_{dt \to 0} \Pr\{T_1 \in [t_1, t_1 + dt) | T_1 \geq t_1, T_2 \geq t_2\}/dt$$

given that the other unit survived to just before t_2.

The cause-specific hazard considered in the context of competing risks (tracking the risk of death due to cause j among survivors to time t) is a special case of the latter, namely the case where $t_1 = t_2$:

$$\lambda_{1|2}(t|T_2 \geq t) = \lim_{dt \to 0} \Pr\{T_1 \in [t, t + dt) | T_1 \geq t, T_2 \geq t\}/dt.$$

Knowledge of the two types of conditional hazards completely determines a joint distribution, see Cox and Oakes (1984, p. 157) for an expression linking the joint density to the conditional hazards.

2.2 Frailty Models

One way to model a joint survival function is to assume the existence of a random effect θ such that given θ, T_1 and T_2 are independent. Depending on the context, θ may represent traits that persist across spells or are common among kin, and which account for the lack of independence.

In symbols, we can write the assumption of conditional independence as

$$S_{12}(t_1, t_2|\theta) = S_1(t_1|\theta) S_2(t_2|\theta),$$

where all survival functions are conditional on θ. Usually the random effect is assumed to act multiplicatively on the hazard, so that

$$S_i(t_i|\theta) = S_{0i}(t_i)^\theta$$

for some baseline survival function $S_{0i}(t)$. Under this assumption the cumulative hazards are

$$\Lambda_i(t_i) = \theta \Lambda_{0i}(t_i)$$

and the individual hazards are

$$\lambda_i(t_i) = \theta \lambda_{0i}(t_i).$$

The *conditional* joint survival function is then

$$\begin{aligned} S_{12}(t_1, t_2|\theta) &= S_{01}(t_1)^\theta S_{02}(t_2)^\theta \\ &= e^{-\theta\Lambda_{01}(t_1)} e^{-\theta\Lambda_{02}(t_2)} \\ &= e^{-\theta(\Lambda_{01}(t_1)+\Lambda_{02}(t_2))}. \end{aligned}$$

This is not very useful because θ is not observed. To obtain the *unconditional* survival function we need to 'integrate out' θ. Suppose that θ has density $g(\theta)$. Then

$$\begin{aligned} S_{12}(t_1, t_2) &= \int_0^\infty S_{12}(t_1, t_2|\theta) g(\theta) d\theta \\ &= \int_0^\infty e^{-\theta(\Lambda_{01}(t_1)+\Lambda_{02}(t_2))} g(\theta) d\theta, \end{aligned}$$

and we recognize this expression as the Laplace transform of $g(\theta)$ evaluated at $s = \Lambda_{01}(t_1) + \Lambda_{02}(t_2)$. Thus

$$S_{12}(t_1, t_2) = \mathcal{L}_g(\Lambda_{01}(t_1) + \Lambda_{02}(t_2)).$$

To make further progress we need to know the distribution of θ.

2.3 Gamma Frailty

Suppose the common or persistent frailty component θ has a gamma distribution with parameters $\alpha = \beta = 1/\sigma^2$. The Laplace transform of the gamma density is $\mathcal{L}(s) = (\beta/(\beta + s))^\alpha$. Using this result,

$$S_{12}(t_1, t_2) = \left(\frac{1}{1 + \sigma^2\Lambda_{01}(t_1) + \sigma^2\Lambda_{02}(t_2)} \right)^{\frac{1}{\sigma^2}}. \tag{1}$$

Actual estimation of this model requires some assumption about the baseline hazards and will be considered in detail further below.

While we are on this subject, it will be useful to write the joint survival function $S_{12}(t_1, t_2)$ under gamma frailty as a function of the marginals $S_i(t_i)$.

We start from the expression for the joint survival and obtain a marginal by setting one of the t's to zero, thus

$$S_1(t_1) = S_{12}(t_1, 0) = \left(\frac{1}{1 + \sigma^2\Lambda_{01}(t_1)} \right)^{\frac{1}{\sigma^2}}.$$

We now use this expression to solve for $\Lambda_{01}(t_1)$, or better still $\sigma^2\Lambda_{01}(t_1)$:

$$S_1(t_1)^{\sigma^2} = \frac{1}{1 + \sigma^2\Lambda_{01}(t_1)},$$

taking reciprocals and subtracting one on both sides gives

$$S_1(t_1)^{-\sigma^2} - 1 = \sigma^2 \Lambda_{01}(t_1),$$

and using this result on Equation 1 we obtain

$$S_{12}(t_1, t_2) = (S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1)^{-\frac{1}{\sigma^2}}. \tag{2}$$

Keep this handy for future reference.

2.4 Non-parametric Frailty

An alternative assumption regarding θ is to treat it as discrete, assuming values $\theta_1, \theta_2, \ldots, \theta_k$ with probabilities $\pi_1, \pi_2, \ldots, \pi_k$, where $\sum \pi_j = 1$.

Laird (1978) and Heckman and Singer (1982, 1984) show that a non-parametric maximum likelihood approach to the estimation of $g(\theta)$ leads precisely to this discrete model. Under the foregoing assumptions, the unconditional survival function is the finite mixture

$$S_{12}(t_1, t_2) = \sum_{j=1}^{k} e^{-\theta_j(\Lambda_{01}(t_1)+\Lambda_{02}(t_2))}\pi_j.$$

Again, estimation of this model requires specifying the baseline hazards $\lambda_{0i}(t)$, and will be considered below.

Other distributional assumptions are possible. Are these models identified? Fortunately yes, as we shall see presently.

2.5 Clayton's Model

Clayton (1978) proposed a continuous bivariate survival model where the two conditional hazards for T_1 given $T_2 = t_2$ and given $T_2 \geq t_2$ are proportional, namely

$$\frac{\lambda_1(t_1|T_2 = t_2)}{\lambda_1(t_1|T_2 \geq t_2)} = 1 + \phi. \tag{3}$$

In words, the risk for unit one at time t_1 given that the other unit failed at time t_2 is $1 + \phi$ times the risk for unit one at time t_1 given that the other unit survived to t_2.

Think of all families with two children whose second child survived to age one, say. Separate those families whose second child died shortly after his or her first birthday. Clearly these families have, on the average, higher risk that the original pool. In fact, the first child in these families is subject to $100\phi\%$ higher risk, at any given age.

Note that the hazard ratio $1 + \phi$ is constant over time. Conditioning on survival (and then death) to age two (instead of age one) in our example would lead to exactly the same hazard ratio.

The remarkable thing about this model is that it is exactly equivalent to a multiplicative frailty model with a gamma-distributed random effect, with $\phi = \sigma^2$.

An important implication of this result is that the model is clearly identified and can be tested, as it has an observable consequence, namely the fact that the ratio of two hazards (which are themselves estimable) is constant over time (something we can verify).

Moreover, this result gives a new interpretation to σ^2, the variance of the random effect. A variance of σ^2 means that children who lost a sibling at age t have a risk $(1 + \sigma^2)$ times the risk of children who had a sibling survive to age t.

Note back on Equation 2 that as $\sigma^2 \to 0$, $S_{12}(t_1, t_2)$ approaches the product of the marginals, as we would expect under independence. As $\sigma^2 \to \infty$, $S_{12}(t_1, t_2)$ approaches $\min\{S_1(t_1), S_2(t_2)\}$, which is known as the Fréchet bound on the maximum possible positive association between two distributions with given marginals.

In other words, the model covers the entire spectrum from independence to maximum possible *positive* association. However, the model cannot account for negative association.

2.6 Oakes's Interpretation

Oakes (1982) showed that ϕ (or σ^2) is closely related to a measure of ordinal association known as Kendall's τ (tau).

Given a bivariate sample $(T_{11}, T_{12}), (T_{21}, T_{22}), \ldots, (T_{n1}, T_{n2})$, Kendall considers all possible pairs of observations. He calls a pair concordant if the first coordinates have the same rank order as the second coordinates. Otherwise a pair is discordant. (For example in husband and wife survival two couples would be concordant if either husband A dies younger than husband B *and* wife A dies younger than wife B, or the A's outlive the corresponding B's.) Kendall's τ is then defined as

$$\tau = \frac{\text{concordant pairs} - \text{discordant pairs}}{\text{number of pairs}}.$$

Unfortunately, when the data are censored we may be unable to calculate this measure. (For example if husband A died younger than husband B, wife

A died, and wife B is still alive but has not yet reached the age at which wife A died, we would not know if the pair is concordant or discordant.)

Oakes has shown that if we restrict the calculation to pairs that can definitely be classified as either concordant or discordant (for example wife B is censored but has already outlived wife A), then the expected value of Kendall's τ under Clayton's model is

$$E(\hat{\tau}) = \frac{\phi}{\phi + 2},$$

which provides further justification for interpreting ϕ (or σ^2) as a measure of ordinal association between kindred lifetimes.

There are no known similar interpretations for frailty distributions other than the gamma. It would be interesting to explore how the ratio of hazards varies over time for other distributions, such as the inverse Gaussian.

3 Multivariate Extensions

The foregoing ideas extend easily to more than two lifetimes and models with observed covariates.

3.1 Notation and Definitions

Consider a set of *clustered* data where

$$\begin{aligned} t_{ij} &= \text{observation time} \\ d_{ij} &= \text{death indicator} \\ x_{ij} &= \text{vector of covariates} \end{aligned}$$

all for the j-th individual in the i-th group (or cluster).

We assume that given x_{ij} and a random effect θ_i the m_i lifetimes in cluster i, say $T_{i1}, T_{i2}, \ldots, T_{im_i}$ are independent.

Thus, the joint distribution of these lifetimes given θ_i is the product of the marginal distributions given θ_i. Under the multiplicative frailty model, the marginal hazards satisfy

$$\lambda_{ij}(t_{ij}|x_{ij}, \theta_i) = \theta_i \lambda_{0ij}(t_{ij}|x_{ij}).$$

Often the covariate effects will be modelled using a proportional hazards model, so that

$$\lambda_{0ij}(t_{ij}|x_{ij}) = \lambda_0(t_{ij}) e^{x'_{ij}\beta}.$$

Combining these two equations we obtain the full model:

$$\lambda_{ij}(t_{ij}|x_{ij},\theta_i) = \theta_i \lambda_0(t_{ij}) e^{x'_{ij}\beta}. \tag{4}$$

In the sections that follow we will often use t_i to denote the vector $(t_{i1}, t_{i2}, \ldots, t_{im_i})'$ of m_i survival times for cluster i, and X_i to denote an m_i by p matrix of covariates with one row for the covariates of each unit in the cluster.

3.2 Gamma Frailty

If the cluster-specific random effects θ_i have independent gamma distributions, then the unconditional survival for the m_i lifetimes in cluster i is

$$S_i(t_i, X_i) = \int_0^\infty \prod_j S_{ij}(t_{ij}|x_{ij},\theta_i) g(\theta_i) d\theta_i,$$

which can be solved easily using the Laplace transform, to give

$$S(t_i, X_i) = \left(\frac{1}{1 + \sigma^2 \Lambda_0(t_{i1}) e^{x'_{i1}\beta} + \ldots + \sigma^2 \Lambda_0(t_{im_i}) e^{x'_{im_i}\beta}} \right)^{\frac{1}{\sigma^2}}.$$

Alternatively, we can write the joint distribution as a function of the marginals. In a direct extension of our earlier result for bivariate distributions,

$$S_i(t_i|X_i) = (S_{i1}(t_{i1}|x_{i1})^{-\sigma^2} + \ldots + S_{im_i}(t_{im_i}|x_{im_i})^{-\sigma^2} - m_i + 1)^{-\frac{1}{\sigma^2}}.$$

Under a proportional hazards model

$$S_{ij}(t_{ij}|x_{ij}) = S_0(t_{ij})^{e^{X'_{ij}\beta}}.$$

3.3 Clayton's Model

Clayton's characterization extends to the multivariate case. We consider the risk for a given individual given the survival status of the others to any given set of ages. Specifically, consider the risk to individual one given that the others have survived to (if $d_{ij} = 0$) or died at (if $d_{ij} = 1$) ages t_{ij}. We will write this conditional hazard as

$$\lambda_1(t_1|t_2, t_3, \ldots, t_m; d_2, d_3, \ldots, d_m),$$

where I have suppressed the group subscript for clarity.

Consider now a similar hazard given that all other members of the group survived to the same ages, which in our notation is

$$\lambda_1(t_1|t_2, t_3, \ldots, t_m; 0, 0, \ldots, 0).$$

Then under a multiplicative frailty model the ratio of these two hazards is constant over time and equals

$$1 + \sigma^2 \sum_j d_j.$$

In the bivariate case this reduces to $1 + \sigma^2$ when the other member of the pair died. In a trivariate case, the risk for the index individual would be $1 + \sigma^2$ if one of the other two died and $1 + 2\sigma^2$ if both of them died.

To clarify this interpretation think of all families with 3 children where the second child survives to age one and the third child survives to age two. These families are subject to the baseline risk. Now consider the subset where the second child died around age one but the third child was alive at age two. These families have higher risk, and their risk is $1 + \sigma^2$ times the baseline. There is another subset where the second child was alive at age one but the third child died around age two. These families also have higher risk and the relative risk factor is also $1 + \sigma^2$. Finally, we have the families where the second child died around age one and the third child died around age two. These families have the highest risk; their relative risk factor is $1 + 2\sigma^2$.

Note that in this model the death of a child is not assumed to have a direct effect on the survival of the remaining siblings. Rather, the death of a child is an indicator that the family has higher than average risk.

3.4 Oakes's Interpretation

Clearly, $\sigma^2/(2 + \sigma^2)$ can still be interpreted as a measure of association between the lifetimes of any two members of the group. As in all models with a single random effect to account for the correlation of three or more r.v.'s, the association between any two members of the group is the same.

Note that this assumption may not always be reasonable. For example kids born closer together in time may face more similar risks than those born farther apart. The model allows independence and maximum positive correlation, but restricts intermediate cases to equal pairwise association.

4 Estimation Using the EM Algorithm

The fact that a multiplicative frailty model would be a standard proportional hazards model–and therefore relatively simple to estimate–if θ_i was observed, suggest immediately the possibility of using the EM algorithm. We now show that this leads to very simple procedures for gamma and for non-parametric frailty. This section follows closely Guo and Rodríguez (1992).

4.1 Gamma Frailty

If θ_i was observed, the likelihood function would depend on the joint distribution of frailty and the survival times T_{ij}. We can write this in terms of the density of θ_i times the conditional distribution of the survival time T_{ij} given θ_i. The contribution of the i-th cluster to the *complete* data log-likelihood would be

$$\log L_i = \log g(\theta_i) + \sum_{j=1}^{m_i} \{d_{ij} \log(\theta_i \lambda_{ij}(t_{ij})) - \theta_i \Lambda_{ij}(t_{ij})\}. \tag{5}$$

The hazard $\lambda_{ij}(t_{ij})$ and cumulative hazard $\Lambda_{ij}(t_{ij})$ will in general depend on the covariates x_{ij} and the parameters β as well as the baseline hazard. I am leaving that implicit to focus on the key aspects of the estimation procedure.

The E-step of the algorithm requires finding the expected value of the complete data log-likelihood, where expectation is taken with respect to the conditional distribution of θ_i given the data. Given the structure of $\log L_i$, this reduces to finding the expected value of θ_i and $\log \theta_i$ given (t_{ij}, d_{ij}) for $j = 1, \ldots, m_i$.

It turns out that this is not hard at all. Direct integration (like we did in the univariate case) shows that if the marginal (or prior) distribution of θ_i is gamma with parameters α and β (usually $\alpha = \beta = 1/\sigma^2$), then the conditional (or posterior) distribution of θ_i given the survival experience of the i-th cluster is also gamma, but with parameters

$$\alpha^* = \alpha + \sum_j d_{ij} \qquad \text{and} \qquad \beta^* = \beta + \sum_j \Lambda_{ij}(t_{ij}).$$

Note that α increases by the total number of deaths and β increases by the total cumulative hazard (or exposure to risk).

The expected value of θ_i given the data is then

$$\mu_i = E(\theta_i) = \frac{\alpha^*}{\beta^*} = \frac{\alpha + \sum_j d_{ij}}{\beta + \sum_j \Lambda_{ij}(t_{ij})}.$$

We could rewrite this expression in terms of σ^2, the variance of frailty, but it turns out to be easier to work with $\alpha(=\beta)$, which may be interpreted as a precision parameter.

The expected value of $\log\theta_i$ when θ_i has a gamma distribution is well known, and in this case turns out to be:

$$\xi_i = E(\log\theta_i) = \Psi(\alpha^*) - log\beta^* = \Psi(\alpha + \sum d_{ij}) - \log(\beta + \sum \Lambda_{ij}(t_{ij})),$$

where Ψ is the digamma function (the first derivative of the log of the gamma function, so $\Psi(x) = \Gamma'(x)/\Gamma(x)$).

Let $\hat{\mu}_i$ and $\hat{\xi}_i$ denote the expected values of θ_i and $\log\theta_i$ evaluated at current parameter estimates. Then the result of the E-step is

$$\begin{aligned} Q_i = \quad & (\alpha-1)\hat{\xi}_i - \alpha\hat{\mu}_i + \alpha\log\alpha - \log\Gamma(\alpha) + \\ & d_i\hat{\xi}_i + \textstyle\sum_j d_{ij}\log\lambda_{ij}(t_{ij}) - \hat{\mu}_i\sum\Lambda_{ij}(t_{ij}). \end{aligned} \tag{6}$$

The first line comes from the density of θ_i and the second line comes from the conditional survival likelihood.

The M-step requires maximizing Q_i w.r.t. $\alpha(=1/\sigma^2)$ and the parameters in $\lambda_{ij}(t_{ij})$. This step breaks neatly into two separate problems.

The part of $Q = \sum Q_i$ involving α is

$$Q_1 = (\alpha-1)\sum\hat{\xi}_i - \alpha\sum\hat{\mu}_i + n\alpha\log\alpha - n\log\Gamma(\alpha),$$

where n is the number of clusters. The first derivative is

$$\frac{\partial Q_1}{\partial\alpha} = \sum(\hat{\xi}_i - \hat{\mu}_i) + n(1 + \log\alpha - \Psi(\alpha)),$$

and the second derivative is

$$\frac{\partial^2 Q_1}{\partial\alpha^2} = n(\frac{1}{\alpha} - \Psi^{(1)}(\alpha)),$$

where $\Psi^{(1)}$ is the trigamma function. This part can be maximized using a Newton-Raphson algorithm. Calculation of the gamma, digamma and trigamma functions can be accomplished using published algorithms. (All three functions are available in R.)

The part of $Q = \sum Q_i$ involving the remaining parameters is exactly the same as the log-likelihood for a standard survival model where $\hat{\mu}_i$ is treated as an extra relative risk. To see this point note that we can add to Q_i the quantity $\sum d_{ij}\log\hat{\mu}_i$, which does not depend on unknown parameters. Then this part becomes

$$Q_2 = \sum_i\sum_j\{d_{ij}\log(\hat{\mu}_i\lambda_{ij}(t_{ij})) - \hat{\mu}_i\Lambda_{ij}(t_{ij})\},$$

which is a standard survival log-likelihood. For example, if we were using a proportional hazards model with a piece-wise constant baseline hazard, Q_2 would be equivalent to a Poisson log-likelihood.

To summarize, the EM algorithm for this problem involves the following steps. Given initial estimates (obtained, for example, by ignoring the multivariate structure of the data):

1 Estimate the expected value of the random effect θ_i and of its logarithm $\log \theta_i$ for each cluster. Call these $\hat{\mu}_i$ and $\hat{\xi}_i$.

2 Obtain new estimates of the parameters by (a) fitting the model using standard univariate procedures but including $\hat{\mu}_i$ as a known relative risk and (b) solving the Newton-Raphson equations for α.

These steps are repeated to convergence. The algorithm is slow, but extremely robust. It is also comparatively easy to implement, because you can take advantage of existing code for the univariate model.

4.2 Non-parametric Frailty

The EM algorithm for non-parametric frailty is even simpler. We assume that a cluster comes from one of K populations representing different levels of frailty $\theta_1, \ldots, \theta_K$. Let π_k denote the probability that the cluster comes from the k-th population (or level of frailty), with $\sum \pi_k = 1$.

We introduce an indicator variable Z_{ik} that takes the value one if the i-th cluster comes from the k-th population (or level of frailty) and zero otherwise. Note that $\Pr\{Z_{ik} = 1\} = \pi_k$. If the Z_{ik} were observed we would maximize the complete data log-likelihood, to which the i-th cluster contributes

$$\log L_i = \sum_{k=1}^{K} \{z_{ik}(\log \pi_k + \log L_{ik})\},$$

where $\log L_{ik}$ denotes the standard log-likelihood given that the level of frailty is θ_k, namely

$$\log L_{ik} = \sum_{j=1}^{n_i} \{d_{ij} \log(\theta_k \lambda_{ij}(t_{ij})) - \theta_k \Lambda_{ij}(t_{ij})\},$$

which of course would depend on some parameters, say β. (Note that for each cluster only one of the K terms in $\log L_i$ is non-zero.)

The E-step requires taking the expected value of $\log L_i$ given the data, which in turn requires the expected value of the indicator variable Z_{ik}. A

fairly straightforward argument shows that if the prior probability that $Z_{ik} = 1$ is π_k, and given this the likelihood of the data is L_{ik}, then the posterior probability that $Z_{ik} = 1$ given the data is

$$\rho_{ik} = E(Z_{ik}|(t_i, d_i)) = \frac{\pi_k L_{ik}}{\sum_{r=1}^{K} \pi_r L_{ir}}.$$

This follow from Bayes theorem or the definition of conditional probabilities. You just have to be careful that for some cases we are talking about the probability of dying at t_{ij} while for others (censored) we need the probability of being alive just before t_{ij}. Note that ρ_{jk} represents the posterior probability that a cluster comes from the k-th population (or level of frailty). Let $\hat{\rho}_{ik}$ denote this posterior probability evaluated at current parameter estimates. The result of the E-step is then

$$Q_i = \sum_{k=1}^{K} \hat{\rho}_{ik} \log \pi_k + \sum_{k=1}^{K} \hat{\rho}_{ik} \log L_{ik}.$$

Note that the second term is just a weighted average of the log-likelihoods given that frailty has value θ_k, with weights given by the posterior probabilities that frailty has value θ_k.

The M-step requires maximizing $Q = \sum Q_i$ w.r.t. the π_k and the parameters in $\log L_{ik}$, namely the θ_k and β. Again, this problem breaks down neatly into two separate problems.

First, maximizing w.r.t. the π_k's gives the explicit solution

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}_{ik},$$

or the average of the posterior probabilities. This follows directly from the multinomial structure of Q_i. (You may take derivatives to verify this result, but remember the restriction $\sum \pi_k = 1$. The easiest thing to do is work with only $k-1$ probabilities and write $\pi_K = 1 \sum_{i=1}^{K-1} \pi_k$.)

Second, maximizing w.r.t. the θ_k's and β is equivalent to maximizing the standard univariate log-likelihood $\log L_{ik}$ with a twist: each observation contributes to each possible level of frailty with weight equal to its posterior probability of coming from that population.

To fix ideas suppose you are fitting a model with two levels of frailty (or two points of support). Then all you have to do is duplicate all observations, introduce a factor coded 1 for the first copy and 2 for the second (this will give θ_1 and θ_2, give weight $\hat{\rho}_{i1}$ to the first copy and $\hat{\rho}_{i2} = 1 - \hat{\rho}_{i1}$ to the second, and fit as usual. Isn't that easy?

These two results apply quite generally to finite mixture models, not just frailty models; for details see the book by Everitt and Hand (1981).

Note: With gamma frailty we assumed $E(\theta_i) = 1$. Introducing a similar restriction in the present context would impose a constraint on the θ_k's and π_k's and would make life more difficult. A much simpler solution is to leave the θ_k's unrestricted and omit a constant from the baseline hazard. After the model is fit, you can calculate the mean frailty as

$$\bar{\theta} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\theta}_k,$$

and then absorb this into the constant. In practice we divide $\hat{\theta}_k$ by $\bar{\theta}$ to make the new mean one and obtain results comparable with gamma frailty.

4.3 Further Notes

4.3.1 Non-Parametric Hazards

All the procedures discussed so far require a parametric model for the baseline hazard, be it exponential, Weibull, log-logistic, or piece-wise exponential survival.

Clearly it would be nice to have a robust method that, like Cox's partial likelihood, could be used to estimate the main parameters of interest, i.e. β and σ^2, without any assumptions about the form of the hazard.

Clayton and Cuzick (1985) have some interesting results along these lines in a procedure that appears to be closely related to the EM algorithm. I recommend their excellent paper and the ensuing discussion.

4.3.2 Baseline Hazards

In our discussion we allowed a different baseline hazard for each member of a cluster, and even a different vector of coefficients β. This makes sense in the study of series of events, such as birth intervals. In other cases, such as studies of siblings, particularly where the events are not distinguishable, it will probably suffice to have a common baseline and common coefficients.

4.3.3 Accelerating the Algorithm

The results of Louis (1984), described in the technical note on the EM algorithm, can be used to

- speed-up the algorithm, and

- obtain standard errors.

See my paper with Guo for details.

4.3.4 The Incomplete Data log-Likelihood

The EM algorithm is simple and stable, and sometimes it is the only way to proceed, particularly if the correct likelihood is hard to obtain or would require numerical integration.

That is *not* the case here. Both with gamma heterogeneity and a finite mixture model the correct incomplete data log-likelihood is tractable. We have already given expressions for the survival function; the hazard follows easily.

Moreover, the first and second derivatives w.r.t. the parameters can be obtained, opening the possibility of using a Newton-Raphson algorithm. This however, requires good starting values. The accelerated EM algorithm is probably more stable and just as fast.

5 Fixed-Effects Models

In the discussion so far we have treated the cluster-specific effect θ_i as *random*: we postulate a distribution and then estimate the parameters of that distribution.

An alternative approach is to treat the θ_i as *fixed* quantities to be estimated, effectively adding one parameter for each cluster.

One difficulty with this approach is that when the number of parameters to be estimated increases with the number of observations, we generally get *inconsistent* estimates, not only for the offending parameters, but also for the other parameters in the model.

There is, however, a way around these difficulties. It is often possible to eliminate the fixed effects θ_i from the likelihood by suitable *conditioning*. Usually one conditions on a statistic (a function of the data) that is minimal sufficient for θ_i (meaning that the likelihood viewed as a function of θ_i depends on the data only through this statistic, which has the same dimensionality as $theta_i$).

In the present context one can construct a *partial likelihood* that eliminates the θ_i. This approach is discussed by Kalbfleisch and Prentice (1980) under the rubric of stratification, and has been advocated among demographers and economists in a series of papers by Ridder and Tunali.

Here are the basis ideas. We assume a multivariate proportional hazards model where the risk to subject j in cluster i is

$$\lambda_{ij}(t_{ij}) = \theta_i \lambda_0(t_{ij}) e^{x'_{ij}\beta},$$

where θ_i is a fixed cluster effect, $\lambda_0(t_{ij}($ is a baseline hazard and $e^{x'_{ij}\beta}$ is a relative risk, as usual.

Next we construct a partial likelihood separately in each cluster. Let $t_{i1} < \ldots < t_{im_i}$ denote the distinct times of death observed in cluster i. Assume no ties, so only one person dies at each t_{ij}, and let R_{ij} denote the risk set in cluster i just before time t_{ij}. The partial likelihood is

$$L_= \prod_{j=1}^{m_i} \frac{\theta_i \lambda_0(t_{ij}) e^{x'_{ij}\beta}}{\sum_{k \in R_{ij}} \theta_i \lambda_0(t_{ij}) e^{x'_{ik}\beta}},$$

and as you may see, this time it is not just the baseline hazard $\lambda_0(t_{ij})$ but also the cluster-specific effect θ_i that cancels out of the likelihood. (In fact, we don't even need to assume the same baseline hazard for each cluster.)

The overall partial likelihood is obtained as the product of the cluster-specific partial likelihoods over all clusters. Ties can be handled using the standard approximations, such as Peto's or Efron's.

Some important points to note about this procedure:

- Clusters with no deaths do not contribute to the likelihood

This is an unfortunate consequence of the fact that θ_i is a fixed unknown parameter. If there are no deaths, it is conceivable that θ_i is zero. This doesn't happen in random-effects models with continuous frailty because the θ_i are supposed to come from a distribution that puts no mass at zero.

- Covariates that are constant within a cluster also drop out from the likelihood

This point is very important. In a study of child mortality, we cannot estimate the effect of mother's education if we use a family-level fixed effect. The reason is that the term $e^{\beta x_i}$ would appear both in the numerator and denominator of the cluster-specific partial likelihood and would therefore cancel out. Another way to think about this is to note that θ_i captures all influences common to members of a cluster, both observed and unobserved. So you can only estimate the effects of child-level covariates.

Moreover, if a covariate happens to have the same value for all children in a family it would also drop out of the likelihood. Imagine a family with

three girls. This family cannot contribute to estimating the effect of sex on mortality. You might think that this family would contribute to estimating the mortality of girls while other families contribute to estimating the mortality of boys, The problem is that the differences between these families could be due to their fixed effects, and have nothing to do with the sex of their children.

Here, then, lies the main advantage and disadvantage of the technique. One often ends up using only a small fraction of the original data, raising the specter that the cases selected for analysis are very different from the rest. On the other hand one may argue that the are precisely the cases that contain information. Only by looking at children within a family who differ on a trait, and such that one dies and the other doesn't, can we be sure that the apparent effect of the trait is not due to unobserved family characteristics.

Fixed-effects models control for both observed and unobserved cluster characteristics; they solve the omitted variables problem at this level, but cannot estimate the effects of included variables. Random-effects models address the problem of intra-cluster correlation, but can only capture the effects of unobserved cluster characteristics that are uncorrelated with observed covariates. They offer no solution to the omitted variables problem, but can estimate the effects of observed variables at all levels.