## Contents

1 Multivariate survival 3
  1.1 Multivariate survival in practice 3
  1.2 Estimation of multivariate survival function 5
  1.3 Regression models of multivariate survival 6
  1.4 Proportional hazards mixed-effects model 7
  1.5 PHMM: inference and computational aspects 9
  1.6 PHMM: testing on the boundary 14
  1.7 PHMM: model selection 18
  1.8 PHMM: model departure and diagnostics 23
  1.9 Partially proportional hazards mixed-effects model 28
Chapter 1

Multivariate survival

1.1 Multivariate survival in practice

Multivariate survival data arise in many situations in practice. According to the correlation structure, we can roughly divide them into serial, symmetric, and other specific correlations. A typical case of serial correlation is recurrent events; in this case, a subject experiences repeated events of the same type. Symmetric correlation structures arise in general clustered data, such as multicenter clinical trials when a center effect is present. In this case, patients accrued from the same trial center can be considered as approximately similarly correlated. Other specific types of correlation can occur in familial or genetic studies, where the members of the same family are correlated according to their specific genetic and common environmental relationships. In the following we give some examples of different types of multivariate survival data.

Recurrent events

Recurrent events are in contrast to what is sometimes called parallel survival times, because the events have a natural time order. These are sometimes events of the same type, such as repeated infections. At other times, they are related to a multi-state process, such as time to disease relapse then to death. In the second case it is assumed that relapse always occurs before death; when this is not true, competing risks may arise, and see the section on competing risks below. Recurrent events have also been used to refer to visit times in longitudinal data collection [ref]. Due to different ways recurrent events data are generated, different models have been proposed with different assumptions. Cook and Lawless (2007) reviewed developments in this area up to very recently. They roughly divided the recurrent events modeling into rate function based, gap time analysis, and intensity based. Both
CHAPTER 1. MULTIVARIATE SURVIVAL

the rate function and the intensity are defined for the counting process \( N(t) \), which records the cumulative number of events occurring over the time interval \([0, t]\). The rate function is the derivative of the mean of \( N(t) \), \( r(t) = dE\{N(t)\}/dt \), while the intensity is the instantaneous probability of an event occurring at \( t \), given the history of the process. The rate function equals the intensity function for a Poisson process.

On the other hand, the gap times refer to time \( T_1 \) from the start to the first event, \( T_2 \) from the first event to the second event, etc. The analysis of gap times sometimes appear similar to the rest of the analysis methods of multivariate survival data that are discussed in this chapter; one cautionary note is on the induced dependent censoring. Induced dependent censoring can happen, for example, if there is a fixed follow-up time \( c \). In this case the censoring time for \( T_2 \) is \( C_2 = c - T_1 \). As \( T_1 \) and \( T_2 \) are likely dependent, so are \( C_2 \) and \( T_2 \).

For the rest of this chapter, we do not focus on recurrent events. There are circumstances where the methods discussed in this chapter can be applied to recurrent events, and we will specify the conditions. We otherwise refer interested readers to Cook and Lawless (2007) and the relevant literature.

Multi-center clinical trials

Multi-center clinical trials are conducted in order to accrue sufficient numbers of research subjects to carry out rigorous comparisons of experimental therapies to standard care. Ideally the subjects accrued from different trial centers can be considered homogeneous, so that accruing from multiple centers simply speeds up the accrual rate, and a relatively large sample size may be achieved in a reasonable time period. However, in some cases the trial centers may consist of different types of hospitals, such as community hospital and university research hospitals. The patient population may be different at these different types of hospitals. So despite tightly written protocols and rigorous monitoring, center effects may exist for the outcomes of interest. This is evident in an Eastern Cooperative Oncology Group (ECOG) clinical trial in advanced non-small cell lung cancer E1582 (Gray, 1995?). The study compared two different chemotherapy, a standard regimen CAV(?) and an alternating one cycle of CAV followed by another cycle of HEM, and so on. As Gray (1995) discovered, due to the complexity of the CAV-HEM treatment, despite the protocol different centers carried out the treatment differently, leading to heterogeneous treatment effects on patient survival when compared with the standard CAV treatment. Vaida and Xu (2000) discovered that the effect of another important predictor of survival, bone metastasis, also differed from center to center. There has not been scientific explanation for the different bone metastasis effects, other than possibly different patient populations at different centers. Ref?? discussed the importance of considering possible center effects in multicenter clinical trials.
Genetic studies

Familial studies have long been conducted to study the genetic causes of disease susceptibility. Clearly in such studies the outcomes of family members are likely correlated due to the common genetic origin. In addition, family members may also share common environments. Correlation structures may arise. For example, the two parents of a family are typically assumed not to share any genetic information, but they share the same living environment. Offspring are assumed to share half their genetic information with each parent. Siblings are assumed to share half(?) the genetic information with each other. Adopted children are assumed to share no genetic information with their adopted family, but they do share the same environment. Among twins, dizygotic twins are assumed to share half their genetic information, while monozygotic twins are assumed to share all their genetic information. In the next section we will elaborate how these different relationships translate to different correlation structures.

Competing risks

Competing risks are a special type of multivariate survival data. In the previous examples, the survival times of all components of the multivariate survival outcome may potentially be observed, subject to noninformative right censoring. In competing risks, often only the earliest component is observed. The other components are right censored by this earliest component. A typical example of competing risks in cancer studies is tumor progression and death without progression. Note that tumor progression is defined by tumor growth of a pre-specified amount, depending on the measurement method (RECIST ref). It is clear that only one of these two possible outcomes can be observed. Challenges arise because the components are most likely correlated, leading to informative censoring. It is known that under informative censoring [early chapter], the different components cannot be identified.

Other types of multivariate survival

There are other types of multivariate survival data. For example, the survival times of the two eyes of the same person; bilateral recurrence of breast cancer, etc.

1.2 Estimation of multivariate survival function

Nonparametric estimation of multivariate survival function is challenging under censorship, typically due to so-called ‘curse of dimensionality’. Different approaches have been investigated in the literature, including nonparametric maximum likelihood for the bivariate survival function (Prentice, 1999) ....
An aspect of the joint survival distribution is the correlation, say, between two time-to-event random variables. For this purpose copula models are useful since the marginal distributions are treated as a nuisance...

1.3 Regression models of multivariate survival

Often we are interested in the relationship between some predictors and survival outcome, for example, treatment assignment. As is the case for correlated data in general, in regression settings there are at least two different ways of modeling. Depending on the application, the statistician may model the multivariate data marginally, or conditionally. The marginal models need not completely specify the correlation structure of the multivariate survival times, and is computationally more straightforward. This also leads to the robustness of marginal modeling, although correctly specified correlation structure may improve the efficiency of estimation and testing. Conditional modeling often refers to conditioning on latent variables, also called random effects. The correlation structure is then derived from the random effects models, as shown below.

For notational purposes, from here on we refer to the independent units of survival times as clusters, and denote them by $i = 1, ..., n$. Examples of clusters from the above are trial centers, families, etc. Within a cluster $i$, we consider the multivariate survival times $(T_{i1}, ..., T_{ik_i})$. Note that for general purposes we allow the cluster size $k_i$ to vary among different clusters. For theoretical purposes conditions might be imposed, such as the $k_i$’s are i.i.d. (Gamst et al., 2009) [ref].

In the following we first focus on proportional hazards regression models. Other types of regression models will be described towards the end of this chapter.

**Marginal survival models**

The best known marginal survival model is the WLW model (ref). It is based on the Cox regression model, and assumes that each component of the multivariate survival outcome marginally follows the proportional hazards regression model, given the covariates. ... [Cai’s work]

In addition to modeling

**Random effects survival models**

As mentioned above, marginal models have the advantage of being robust and, in addition, computationally straightforward. However, in some applications, the correlation structure itself is of interest, and in other applications, the magnitude of the realized random effects is of interest. For these applications, the random effects survival model is a powerful tool.
1.4. PROPORTIONAL HAZARDS MIXED-EFFECTS MODEL

As is the case for marginal models, and for survival models in general, we may model the regression via the hazard function, or via the survival random variable directly. The extension of the Cox model to incorporating random effects is given by

1.4 Proportional hazards mixed-effects model

The proportional hazards mixed-effects model (PHMM) was initially proposed in Ripatti and Palmgren (2000) and Vaida and Xu (2000), for handling correlated time-to-events data. The general framework of the model draws parallel with the linear, non-linear and generalized linear mixed effects models (LMM, NLMM and GLMM), in that an arbitrary design matrix is allowed for the random effects. The framework encompasses the frailty models previously studied, which also include the hierarchal random effects Cox models (see for example, Yau (2001) and Ma et al. (2003)). For an extensive summary of work done on the frailty models, see Hougaard (2000).

For clustered data let $i = 1, ..., m$ denote the clusters, and $j = 1, ..., n_i$ denote the observations from a cluster. The observed data for individual $ij$ is $(X_{ij}, \delta_{ij}, Z_{ij})$, where $X_{ij} = \min(T_{ij}, C_{ij})$, $\delta_{ij} = I(T_{ij} \leq C_{ij})$, $T_{ij}$ is the potential failure time and $C_{ij}$ is the potential censoring time, $Z_{ij}$ is a vector of covariates. Also let $Y_{ij}(t) = I(T_{ij} > t, C_{ij} \geq t)$ be the ‘at risk’ indicator, $N_{ij}(t) = I(T_{ij} \leq t, T_{ij} \leq C_{ij})$ be the counting process, and $N = \sum_{i=1}^{m} n_i$ be the total sample size. The proportional hazards mixed-effects model can be written as

$$\lambda_{ij}(t) = \lambda_0(t) \exp\{\beta'Z_{ij} + b'_iW_{ij}\},$$

(1.1)

where $\lambda_{ij}(\cdot)$ is the hazard function for individual $ij$, $\lambda_0(\cdot)$ is the baseline hazard, $\beta$ and $b_i$ are vectors of fixed and random effects, respectively, and $W_{ij}$ is usually a sub-vector of $Z_{ij}$ corresponding to those covariates that have random effects. $W$ may also include a 1 if there is a random effect on the baseline hazard itself. When $W = 1$, (1.1) becomes the shared frailty model.

In general the inclusion of the $b'W$ term in the model represents cluster by covariate interactions, such as a treatment by center interaction in multi-center clinical trials. As an example, if a single covariate $Z = 0$ or 1 indicates one of two treatment assignments, then the following specification of model (1.1) can be used to capture center-specific treatment effects:

$$\lambda_{ij}(t) = \lambda_0(t) \exp\{b_iZ_{ij}\}.$$ 

In the above $\beta + b_i$ has the same interpretation as the regression coefficient (fixed effect) when using center as a categorical variable in a classic proportional hazards
model with no random effects; however, the estimation here would be different, as will be explained below.

In contrast to the fixed effects modeling mentioned above, we assume that the $b_i$'s are randomly sampled from a distribution. This serves the purpose of ‘borrowing strength’ and reducing the number of parameters that need to be estimated (Efron and Morris, 1975; Morris, 1983). Since random effects are allowed for arbitrary covariates in the model, the distribution of the random effects should be scale-invariant, like normal or $t$-distribution. Note that the commonly used gamma distribution under frailty models is not scale-invariant; a linear change of scale in the covariates will result in a random effect distribution that no longer belongs to the gamma family. As we generally assume that a covariate with a random effect also has a fixed effect, we may set the random effects to have mean zero. In the following we assume that the $b_i$'s are independent and identically distributed ($i.i.d.$) according to $N(0, \Sigma)$.

Model (1.1) is a proportional hazards model conditional on the random effects. When the random effects are integrated out, except for special cases the resulting marginal model will not have proportional hazards; this is well-known from the literature of missing covariates under the proportional hazards model [add ref or details]. From here we see that the random effects and the marginal approaches involve different assumptions about the proportional hazards, and the two cannot be true at the same time.

Applications

There are at least three types of applications of the PHMM. In the first type of applications the main quantities of interest are the fixed effects, while the correlation in the data needs to be taken into account to ensure the consistency of estimation. In the second type of applications the magnitude of the random effects is of primary interest. Both types of applications were illustrated in Vaida and Xu (2000) using a multi-center clinical trial for lung cancer, where the treatment effect was found to vary significantly from center to center. Sylvester et al. (2002), Glidden and Vittinghoff (2004) and Murray et al. (2004) also argued favorably for using random effects survival models to analyze multi-center trial data. The third type of applications often occur in genetic research where the variance components are the primary focus; such applications can be found in Ripatti et al. (2003) and Liu et al. (2004ab, 2005).
1.5  PHMM: inference and computational aspects

Conditional on the random effects, the observations from the same cluster are assumed to be independent. The clusters are also assumed to be i.i.d. A more relaxed assumption can also be made; see Gamst et al. (2009). In addition, we assume that conditional on the covariates, the censoring time is independent of the failure time and the random effects.

Nonparametric maximum likelihood estimator

Inference under the PHMM is carried out using the likelihood. Each observation can be written $y_{ij} = (X_{ij}, \delta_{ij}, Z_{ij}, W_{ij})$, and the inference is conditional on the covariates, as is often the case in regression settings. Let $y_i = (y_{i1} \ldots y_{in_i})$, i.e. the data for cluster $i$. Conditional on the random effects, the (full) log-likelihood for the right censored data in cluster $i$ is

$$l_i(\beta, \lambda_0; y_i | b_i) = \sum_{j=1}^{n_i} \{ \delta_{ij} \log \lambda_0(X_{ij}) + \delta_{ij} (\beta'Z_{ij} + b_i'W_{ij}) - \Lambda_0(X_{ij}) e^{\beta'Z_{ij} + b_i'W_{ij}} \},$$

(1.2)

where $\Lambda_0(t) = \int_0^t \lambda_0(s) \, ds$ is the cumulative baseline hazard. To keep the notation simple the above expression assumes no ties in the observed failure times, otherwise $X_{ij}$ should be replaced by the unique time points. Denote $\theta = (\beta, \lambda_0, \Sigma)$; strictly speaking in place of $\Sigma$ should be the unknown parameters in that covariance matrix. The likelihood of the observed data is then

$$L(\theta) = \prod_{i=1}^n \int L(\theta|b_i) p_{\Sigma}(b_i) \, db_i = \prod_{i=1}^n \int \exp(l_i) p_{\Sigma}(b_i) \, db_i,$$

(1.3)

where $L(\theta|b) = \prod_i \exp(l_i)$ is the likelihood conditional on the random effects $b = (b_1, \ldots, b_n)$. Usually no closed-form expression is available for $L(\theta)$ and its calculation involves $d$-dimensional integration.

Discretize the baseline hazard. previous chap?

The likelihood (1.3) above is a full likelihood, because it involves the unknown baseline hazard function. Under the classic Cox model, the partial likelihood is a profile likelihood after profiling the baseline hazard out from the full likelihood (previous chap??). In the presence of the random effects, the baseline hazard function cannot be profiled out from (1.3) algebraically, so one does not arrive at a partial likelihood. On the other hand, conditional on the random effects the conditional partial likelihood is

$$\prod_{i=1}^m \prod_{j=1}^{n_i} \frac{Y_{ij}(X_{ij}) \exp(\beta'Z_{ij} + b_i'W_{ij})}{\sum_{k=1}^n \sum_{l=1}^{n_{il}} Y_{kl}(X_{ij}) \exp(\beta'Z_{kl} + b_k'W_{kl})}.$$  

(1.4)
CHAPTER 1. MULTIVARIATE SURVIVAL

The above partial likelihood has been used in other estimators discussed at the end of this section.

The nonparametric maximum likelihood is consistent, asymptotically normal and asymptotically efficient under model (1.1). The main ideas of the proofs are similar to Murphy (1994, 1995) and Parner (1998) for gamma frailty models, plus an identifiability argument from Zeng et al. (2005) for mixed-effects models. Details are given in Gamst et al. (2009).

**EM algorithm**

Vaida and Xu (2000) developed the nonparametric maximum likelihood estimate (NPMLE) of the parameters under the PHMM using a Monte Carlo EM algorithm.

After discretizing the baseline hazard function to point masses on the observed failure times, the EM algorithm operates on finite number of unknown parameters given the data, and the non-descending property of the EM algorithm carried over. That is, the likelihood evaluated at each iteration is non-decreasing. In the following we describe the details of the E- and the M-step.

In the E-step we compute the conditional expectation of the log-likelihood of the so-called augmented data $(y, b)$ given the observed data $y$ and the current parameter value $\theta^*$. Let

$$Q(\theta) = E_{\theta^*}\{l(\theta; y, b) \mid y\} = E_{\theta^*}\{l(\beta, \lambda_0; y \mid b) \mid y\} + E_{\theta^*}\{\log p(y \mid b) \mid y\} = Q_1(\beta, \lambda_0) + Q_2(\Sigma).$$

(1.5)

Note that in the above $Q(\theta)$ separates into two terms: the first term involves only $\beta$ and $\lambda_0$, and is denoted by $Q_1(\beta, \lambda_0)$; the second term involves only $\Sigma$, and is denoted by $Q_2(\Sigma)$. Straightforward algebra shows that

$$Q_1(\beta, \lambda_0) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \delta_{ij} \{\log \lambda_0(t_{ij}) + \beta'Z_{ij} + E(b_i)'W_{ij}\} - \Lambda_0(t_{ij}) \exp \left\{\beta'Z_{ij} + \log E\left(e^{b_i'W_{ij}}\right)\right\}.$$  

(1.6)

In the above the expectations are taken conditionally on $y_i$ and under the current parameter value $\theta^*$, which for simplicity are omitted from notation; also we write $\log E(e^{b_i'W_{ij}})$ as an ‘offset’ in the linear predictor of the Cox regression, which enables the use of common software in the M-step below. The expression of $Q_2(\Sigma)$ depends on the structure of $\Sigma$; for example, if it is diagonal with elements $\sigma^2_1, ..., \sigma^2_d$,

$$Q_2(\Sigma) = -\frac{dn}{2} \log(2\pi) - \frac{1}{2} \sum_{k=1}^{d} \left\{2n \log \sigma_k + \frac{1}{\sigma_k^2} \sum_{i=1}^{n} E(b_{ik}^2)\right\}.$$  

(1.7)
From the above it is clear that the conditional expectations we need in the E-step are $E(b_i)$, $E(e^{b_i'W_{ij}})$, and $E(b_{ik}^2)$ if $\Sigma$ is diagonal. It turns out that $E(b_i)$ is not used in the EM iteration, but it gives the estimates (prediction) of the random effects at the convergence of the EM. Since the random effects are normally distributed, and the Cox model likelihood is basically a Poisson one (ref), they are not conjugates and the expectations do not have closed forms. On the other hand, the gamma distribution is conjugate with the Poisson distribution, and therefore in gamma frailty models they have closed forms; in fact, the observed data likelihood has a closed form in that case. In general the expectations are approximated using numerical methods (quadrature for example) or Monte Carlo simulation. Gibbs sampling is implemented for this purpose in the R package ‘phmm’.

The separation of $Q(\theta)$ in (1.5) means that we can maximize $Q_1$ and $Q_2$ with respect to $\beta$ and $\lambda_0$, and $\Sigma$ separately. Note that $E(b_i'W_{ij})$ and $E(e^{b_i'W_{ij}})$ are constants in maximizing $Q_1$ over $\beta$ and $\lambda_0$, and that the expression of $Q_1$ is very similar to the full log-likelihood under the classic Cox model. Following the derivation of Chapter ??, we can show that given $\beta$ the value of $\lambda_0$ that maximizes $Q_1$ is $\hat{\lambda}_0$ with point masses $\hat{\lambda}_1, \ldots, \hat{\lambda}_s$ on the $s$ distinct observed event times, with

$$\hat{\lambda}_g = \frac{1}{\sum_{t_{kl} \geq t_g} \exp\{\beta'Z_{kl} + \log E(e^{b_k'W_{kl}})\}}$$

for $g = 1, \ldots, s$. This is the usual Breslow’s estimate of baseline hazard with ‘offset’ terms $\log E(e^{b_i'W_{ij}})$. Substituting (1.8) back into $Q_1$ gives (apart from a constant)

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \delta_{ij} \left[ \beta'Z_{ij} - \log \sum_{t_{kl} \geq t_{ij}} \exp\{\beta'Z_{kl} + \log E(e^{b_k'W_{kl}})\} \right].$$

The above is the same as the log partial likelihood under the classic Cox model, again with offsets $\log E(e^{b_i'W_{ij}})$. This way we can use the standard software for fitting the classic Cox model in the M-steps for updating $\beta$ and $\lambda_0$.

To maximize $Q_2$, notice that it is the log likelihood of $n$ independent normally distributed mean zero random variables $b_i$, where the sufficient statistics are replaced with their conditional expectations. This leads to the solution as the usual maximum likelihood estimator with the sufficient statistics replaced by their conditional expectations. For example, if $\Sigma$ is diagonal as before, then

$$\hat{\sigma}_k^2 = \frac{1}{n} \sum_{i=1}^{n} E(b_{ik}^2)$$

for $k = 1, \ldots, d$. If $\Sigma$ is unconstrained, then

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} E(b_ib_i')$$

(1.10)
CHAPTER 1. MULTIVARIATE SURVIVAL

The EM iteration typically starts with the classic Cox model fit with no random effects as the initial value for $\beta$ and $\lambda_0$, and $\Sigma$ may be taken as the identity matrix. Sometimes, if the variances in $\Sigma$ are believed to be very small in magnitude, such as in certain genetic studies, it might speed up convergence using smaller initial values. Declaration of zero variances of the random effects may be addressed using the likelihood ratio test or model selection methods below ???. Automated stopping rules have been proposed in the literature (Vaida and Xu, 2000; Ripatti et al., 2002, and references therein). In our experience, however, it has been the most informative to visually inspect the convergence of the parameter values. Plots of all the parameter values over the iterations are available in the R package ‘phmm’.

Variance estimate

Following the EM algorithm we can compute the Fisher information of the NPMLE $\hat{\theta}$ using Louis’ (1982) formula, and the explicit formulas were given in Vaida and Xu (2000). Gamst et al. (2009) showed that this provides a consistent estimate of the asymptotic variance of $\hat{\theta}$. Since the dimension of the Fisher information matrix is of the same order as the number of failure events, its inversion can be difficult if the number (dimension) is too large. Often we are interested in the estimation of $\beta$ and $\Sigma$ only, while $\lambda_0$ is considered a nuisance. In this case the profile likelihood discussed below can be used.

Example

In this chapter we use the Eastern Cooperative Oncology Group (ECOG) 1582 study which is a randomized multi-center clinical trial in advanced stage lung cancer, to illustrate the various methodology that we describe. There were 31 institutions with a total of 579 patients, and the endpoint is overall survival. As mentioned earlier, this is an example where the clusters are the institutions, and the random effects represent cluster effects or cluster by covariate interactions. The covariates that are known to impact overall survival are: treatment, presence or absence of bone metastases, presence or absence of liver metastases, performance status at entry, and whether there was weight loss prior to entry. Table 1.5 gives the model fits for the Cox model with all five covariates and no random effects, with a single random treatment effect, and with random treatment and bone metastases effects. Any additional random effects added to the model turned out to have their variances converging to zero during the EM, and therefore is not included.
1.5. PHMM: INFERENCE AND COMPUTATIONAL ASPECTS

Table 1.1: NPMLE for E1582 lung cancer data

<table>
<thead>
<tr>
<th></th>
<th>(d = 0)</th>
<th>(d = 1)</th>
<th>(d = 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>(-0.25 (0.09))</td>
<td>(-0.25 (0.10))</td>
<td>(-0.25 (0.12))</td>
</tr>
<tr>
<td>bone</td>
<td>0.22 (0.09)</td>
<td>0.21 (0.10)</td>
<td>0.23 (0.14)</td>
</tr>
<tr>
<td>liver</td>
<td>0.43 (0.09)</td>
<td>0.42 (0.09)</td>
<td>0.39 (0.09)</td>
</tr>
<tr>
<td>ps</td>
<td>(-0.60 (0.10))</td>
<td>(-0.64 (0.11))</td>
<td>(-0.65 (0.13))</td>
</tr>
<tr>
<td>wt loss</td>
<td>0.20 (0.09)</td>
<td>0.22 (0.09)</td>
<td>0.21 (0.09)</td>
</tr>
</tbody>
</table>

\begin{align*}
\text{treatment} & \quad \sigma \\
\text{bone} & \quad - \\
\end{align*}

Note: \(d\) is the number of random effects, \(\sigma^2\) is the variance of each random effects (assume independent when there are two); in (·) are the standard errors.

Computing the likelihood

Although the parameter estimation via EM algorithms has avoided computing the likelihood (1.3), it is sometimes useful to evaluate it at given parameter values, including at the NPMLE. Notice that this is different from having to evaluate it at many different parameter values, and search for the MLE for example; the difficulty there was illustrated in Pinheiro and Bates (1995).

Xu et al. (2009) compared three general methods to evaluate the likelihood: 1) Laplace approximation, 2) reciprocal importance sampling, and 3) bridge sampling. Laplace approximation replaces the integrand in (1.3) by a normal density for each cluster, and has a closed-form expression. It is known to be relatively accurate if the cluster size \(n_i\) is large. In Xu et al. (2009) it agreed with the two sampling methods when \(n_i = 20\). The two sampling methods are, of course, more computationally intense. They turned out to agree with each other for all scenarios considered in that paper. Since the bridge method involves an additional layer of sampling, the reciprocal importance sampling is recommended.

Other estimators and comparison with NPMLE

Ripatti and Palmgren (2000) considered the penalized partial likelihood (PPL) using the conditional partial likelihood given in (1.4). Their approach is similar to the penalized quasi-likelihood (PQL) of Breslow and Clayton (1993). ... The PPL is faster to compute than the NPMLE, although its theoretical properties have not been established; in particular, the estimation of its variance remains a challenge.
Gamst et al. (2009) carried out simulation studies to compare the NPMLE and the PPL. They discovered that the NPMLE is numerically stable and accurate in general, while the PPL estimate is accurate only when the cluster sizes are reasonably large. Also as mentioned above, they discovered that the variance of the PPL is not well estimated in general.

Another estimator that was considered in Gamst et al. (2009) was an approximate NPMLE proposed by Cortiñas Abrahantes and Burzykowski (2005), where the MCMC in the E-step was replaced by the Laplace approximation. Like the PPL, this approximate NPMLE worked fine when the cluster sizes are reasonably large, and becomes inaccurate when they become smaller. Cortiñas-Abrahantes et al. (2007)??

With today’s computation capability, MCMC has been widely used in applications. To us it is clear that the NPMLE should be the frequentist estimator of choice.

Bayesian approaches has also been proposed using the conditional partial likelihood (1.4) (Sargent, 1998). It is beyond this text to discuss the fact that the partial likelihood may not be a bona fide likelihood. Bayesian model selection and model average also provides ways of estimation under the PHMM; see the model selection section below.

1.6 PHMM: testing on the boundary

Hypothesis testing under the PHMM can be carried out following the NPMLE-based inference, if the null hypothesis lies within the parameter space. A particular type of null hypotheses of interest, however, lie on the boundary of the parameter space; that is, to test whether the variances of one or more random effects are zero.

Chernoff (1954) first considered the distribution of the likelihood ratio test statistic when under the null hypothesis the parameter value is a boundary point of a cone; it showed that its asymptotic distribution is the same as one based on a single observation from a multivariate normal distribution. Self and Liang (1987) used the Chernoff (1954) results to establish the $\sqrt{n}$-consistency of the maximum likelihood estimator when the true parameter value $\theta_0$ is the vertex of a cone $C$ by which the parameter space $\Omega$ can be approximated; that is, $\inf_{x \in C} \| x - y \| = o(\| y - \theta_0 \|)$ for all $y \in \Omega$, and $\inf_{y \in \Omega} \| x - y \| = o(\| x - \theta_0 \|)$ for all $x \in C$. Self and Liang (1987) further showed that the multivariate normal representation can be used to derive the asymptotic distribution of the likelihood ratio test when the null hypotheses lie on the boundary of the parameter space. These results were applied by Stram and Lee (1994, 1995) to linear mixed effects models.

The same derivation above carries over to the semiparametric PHMM, mainly due to the asymptotic quadratic expansion of the log profile likelihood, initially established by Murphy and van der Vaart (2000) for the frailty model. In fact,
much of the classical likelihood theory is derived from the asymptotic quadratic form of the log likelihood; and when the quadratic form carries over to the log profile likelihood, the corresponding theory does, too. Such a quadratic expansion was not directly established under the PHMM like in Murphy and van der Vaart (2000); however, asymptotic normality of the NPMLE implies that the likelihood surface is asymptotically quadratic near the true parameter value, which in turn implies that the same holds for the profile likelihood. Maple et al. (2002) also verified empirically that the contours of the profile likelihood under the multinormal PHMM are elliptic.

\textit{The profile likelihood}

Denote $\phi = (\beta, \Sigma)$ the parameter of interest under the PHMM. The log profile likelihood function for $\phi$, with the nuisance parameter $\lambda_0$ ‘profiled out’, is

$$\text{pl}(\phi) = \sup_{\lambda_0} l(\phi, \lambda_0).$$

(1.11)

Under suitable conditions the log profile likelihood behaves as a quadratic function asymptotically; that is, for any random sequence $\phi_n$ such that $\|\phi_n - \phi_0\| = O_p(1/\sqrt{n})$, where $\phi_0$ is the true parameter value,

$$\frac{1}{n} \{\text{pl}(\phi_n) - \text{pl}(\phi_0)\} = (\phi_n - \phi_0)'A - \frac{1}{2}(\phi_n - \phi_0)'I(\phi_n - \phi_0) + o_p\left(\frac{1}{n}\right).$$

(1.12)

Here $A = \sum_1^n s(y_i)/n$, $s$ is the efficient score for $\phi$, i.e., the ordinary observed score function minus its orthogonal projection onto the closed linear span of the score functions for the nuisance parameter $\lambda$; and $I$, its covariance matrix, is the efficient Fisher information matrix (Murphy and van der Vaart, 2000; Severini and Wong, 1992).

The profile likelihood can be used to define a semiparametric likelihood ratio. For two nested models, let $\Theta$ be the parameter space under the larger model and $\Theta_0$ the parameter space under the smaller model or, equivalently, under the null hypothesis $H_0$. Assume that $H_0$ places no restrictions on the nuisance parameter $\lambda$. Write $L$ for the likelihood, and let

$$LR = \frac{\sup_{\Theta_0} L(\phi, \lambda)}{\sup_{\Theta} L(\phi, \lambda)}.$$  

(1.13)

Then $LR$ is the ratio of the maximized likelihoods under the two models, and can be viewed as the ratio of the maximized profile likelihoods, with the nuisance parameter $\lambda$ ‘profiled out’. So

$$-2 \log LR = -2\left\{\sup_{\phi} \text{pl}(\phi) - \sup_{\phi_0} \text{pl}(\phi_0)\right\},$$

(1.14)
where \( \Phi_0 \) and \( \Phi \) are the corresponding parameter spaces for \( \phi \) under the two models. Murphy and van der Vaart (2000) showed that as result of the quadratic expansion (1.12), when \( \phi_0 \) lies in the interior of the parameter space, the maximum likelihood estimator of \( \phi \) is asymptotically normal, and the profile likelihood ratio test for \( H_0 : \phi = \phi_0 \) has asymptotically a chi-squared distribution with degrees of freedom equal to the dimension of \( \phi \) under the null hypothesis \( H_0 \).

Apply to the PHMM

Using (1.11) Xu et al. (2009) obtained the \( \sqrt{n} \)-consistency of the maximum (profile) likelihood estimator under the PHMM when \( \phi_0 \) is on the boundary of its parameter space \( \Phi \). The right hand side of (1.12) can be written as

\[
\frac{1}{2} A' I^{-1} A - \frac{1}{2} \{ z_n - (\phi_n - \phi_0) \} I \{ z_n - (\phi_n - \phi_0) \} + o_p \left( \frac{1}{n} \right),
\]

where \( z_n = I^{-1} A \). Therefore the same representation of the asymptotic distribution of \(-2 \log LR\) as those of Chernoff (1954) and Self and Liang (1987) is obtained, which can then be used to calculate the null distribution of the likelihood ratio statistics. Specifically, assume that \( \Phi \) and \( \Phi_0 \) are regular enough to be approximated by cones with vertices at \( \phi_0 \). Let \( Z \) be a random variable with a multivariate Gaussian distribution with mean \( \phi \) and covariance matrix \( I^{-1}(\phi_0) \), and let \( C_\Phi \) and \( C_{\Phi_0} \) be non-empty cones approximating \( \Phi \) and \( \Phi_0 \) at \( \phi_0 \), respectively. Then the asymptotic distribution of the likelihood ratio statistic, \(-2 \log LR\), is the same as the distribution of the likelihood ratio test of \( \phi \in C_{\Phi_0} \) versus \( \phi \in C_\Phi \) based on a single realization of \( Z \) when \( \phi = \phi_0 \).

Using the above representation one can derive the asymptotic null distribution of the profile likelihood ratio statistic under the PHMM, just like Stram and Lee (1994, 1995) did for linear mixed effects models (they had an error for Case 3 below). Xu et al. (2009) listed some of the cases which are the most likely to be encountered in practice. In the following \( d \) is the dimension of \( b \).

Case 1: \( d = q + 1 \) and

\[
\Sigma = \begin{pmatrix}
\Sigma_{11} & \sigma_{12} \\
\sigma_{12} & \sigma_{22}
\end{pmatrix},
\]

where \( \Sigma_{11} \) is \( q \times q \) and \( q \geq 0 \). The asymptotic null distribution of \(-2 \log LR\) for testing \( H_0 : \sigma_{22} = 0 \) against \( \Sigma \) positive semidefinite is \( (\chi_q^2 + \chi_{q+1}^2)/2 \). When \( q = 0 \), the distribution is a 50:50 mixture of a point mass at 0 and \( \chi_1^2 \); note that in this case the maximum likelihood estimator of the variance components has a positive probability of being zero.

Case 2: Same as in Case 1, but the test also includes a \( r \)-dimensional subvector of fixed effects, \( \beta_1 \), i.e., \( H_0 : \sigma_{22} = 0, \sigma_{12} = 0, \beta_1 = 0 \) against \( \Sigma \) positive semidefinite and \( \beta_1 \neq 0 \). The asymptotic distribution of \(-2 \log LR\) is \( (\chi_{q+r} + \chi_{q+r+1})/2 \).
1.6. PHMM: TESTING ON THE BOUNDARY

Case 3: \( d = q + k \) and

\[
\Sigma = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma'_{12} & \Sigma_{22}
\end{pmatrix},
\]

where \( \Sigma_{11} \) is \( q \times q \) and \( \Sigma_{22} \) is \( k \times k \). The asymptotic null distribution of \(-2 \log LR\) for testing \( H_0: \Sigma_{22} = 0 \) against \( \Sigma \) positive semidefinite is a mixture of \( \chi^2 \) distributions with degrees of freedom \( s, s+1, \ldots, s+k \), where \( s = kq + k(k-1)/2 \). The mixing probabilities are not directly available in general, but can be obtained using simulation. If, in addition, the condition \( \beta_1 = 0 \) is part of the null hypothesis, then the asymptotic distribution of \(-2 \log LR\) is a \( \chi^2 \) mixture with degrees of freedom \( s+r, \ldots, s+r+k \).

Case 4: Another model of interest has \( \Sigma_{12} = 0 \) and \( \Sigma_{22} \) is diagonal. Similarly to Case 3, the asymptotic null distribution for testing \( \Sigma_{22} = 0 \) is a \( \chi^2 \) mixture with degrees of freedom 0 through \( k \).

The above results are asymptotic in the sense that the number of clusters, \( n \), should go to infinity. For small \( n \), Crainiceanu and Ruppert (2004) derived the mixing probabilities for balanced linear one-way ANOVA with a single variance component, and when the cluster sizes are large. Typically the mass at zero is greater than 0.5. Limited simulation in Xu et al. (2009) showed that their mixing probabilities seemed to hold approximately, when the cluster sizes are large.

Computation of the likelihood ratio statistic only requires the profile likelihood at the NPMLE, which is the same as the full likelihood at the NPMLE. The methods for computing the likelihood described in the previous section can be applied.

Example

Continuing with the E1582 lung cancer data, we compute the log-likelihood at the NPMLE using the methods described in the previous section, for the models considered in Table 1.5. All three methods give comparable values in this case, with reciprocal importance sampling (RIS) and bridge sampling differing by no more than 0.025, and Laplace differing from these two by no more than 0.1. Negative twice of the log-likelihood at the NPMLE given by RIS for models with \( d = 0, 1 \) and 2 are in Table 1.2. The likelihood ratio statistics for testing \( d = 0 \) versus \( d = 1 \), and \( d = 1 \) versus \( d = 2 \) are then 4 and 6.25, respectively. Notice that both tests belong to Case 1 above, and have an asymptotic null distribution of \((\chi^2_0 + \chi^2_1)/2\). The critical value for 0.05 significance level is 2.71. It is clear that both tests reject the null hypothesis of no treatment or bone metastases random effect. A test of \( d = 0 \) versus \( d = 2 \) has an asymptotic null distribution that is a mixture of \( \chi^2_0, \chi^2_1 \) and \( \chi^2_2 \) (Case 4). The 0.95 quantile of \( \chi^2_3 \) is 5.99, hence it is clear that with a likelihood ratio statistic of 10.25 we reject \( d = 0 \) in favor of \( d = 2 \).
1.7 PHMM: model selection

There are many ways of model selection. Hypothesis testing can be used for this purpose, measures such as R-squared are also used for this purpose. Here we concentrate on information criteria, and in particular, the Akaike information criteria (AIC). Other information criteria, such as the Bayesian information criteria (BIC), are briefly discussed towards the end of this section. For relationship between these different criteria, we refer to ....

It has been recognized that the method or criterion used for model selection depends, among other things, the focus of the reference. This has been made most explicit by the Focused information criteria (Claeskens and Hjort, 2003; Hjort and Claeskens, 2006, FIC). In the context of mixed effects models, Vaida and Blanchard (2005) distinguished between the marginal focus and the conditional focus. For clustered data, the conditional focus is on the cluster level inference, where any future observation takes place in the same clusters as the observed data. On the other hands, the marginal focus is on the population level inference, where a future observation takes place in a future cluster, assumed to be sampled from the same population as the observed clusters. Both types of focus can be relevant in practice. The marginal focus has typically been considered in the classic (frequentist) literature, where the only unknown quantities are the population parameters. The conditional focus, however, also concerns the unknown magnitude of the random effects, which are often of interest in applications such as small area estimation. We also provide in example below how it can be relevant in biomedical research.

Under the PHMM, both focuses have the baseline hazard function as a nuisance parameter. Therefore the developments here follows the profile likelihood arguments above.

Profile AIC

The Akaike information is defined based on the Kullback-Leibler (KL) information, which is roughly a ‘distance’ (asymmetric nonetheless) from the true distribution that has generated the data \( Y \), to a member of the model family \( g_\theta \): \( \text{KL}(f, g_\theta) = E_f \{ \log f(Y) - \log g_\theta(Y) \}. \) In the presence of a nuisance parameter say \( \lambda \), Xu et al. (2009) argue that the relevant distance is that between \( f \) and the subfamily of models defined by a given value of \( \phi \), the parameter of interest, while allowing \( \lambda \) to vary freely. This distance is then \( \min_{\lambda \in \Lambda} \text{KL}(f, g_{\phi,\lambda}) \), where \( g \) is a member of the model family, and \( \Lambda \) is the parameter space for \( \lambda \). Suppose that the above minimum is attained at some \( \lambda = \tilde{\lambda}(\phi) \) for each \( \phi, \tilde{\lambda}(\phi) \) is in fact a least favorable curve under smoothness conditions (Severini and Wong, 1992; Fan and Wong, 2000). Write \( g_\phi = g(\cdot|\phi, \tilde{\lambda}(\phi)) \); ignoring the constant term \( E\{\log f(Y)\} \) in KL we have \( E\{\log g_\phi(Y)\} = \max_\lambda E\{\log g_{\phi,\lambda}(Y)\} \). Therefore \( g_\phi \) is the theoretical equivalent of
the profile likelihood.

Parallel to the definition of the classic Akaike information, we want to minimize the above distance over \( \phi \). Suppose that the minimum is attained at \( \phi_0 \), i.e. \( E\{\log g_{\phi_0}(Y)\} = \max_{\phi} E\{\log g_{\phi}(Y)\} \). Then \( g_{\phi_0} \) is the best approximation to \( f \) within the family of models; when the model is correct, we have \( f = g_{\phi_0} \). In practice \( \phi_0 \) is estimated by \( \hat{\phi} \) which maximizes the profile likelihood, i.e. the MLE.

The Akaike information, called the profile AI (pAI) in this case, is defined as the predictive value of the profile likelihood \( pl(Y^*|\hat{\phi}) \) for new data \( y^* \), independent of but from the same distribution as the original data \( y \):

\[
\text{pAI} = -2E_{f(y)}E_{f(y^*)}\{pl(y^*|\hat{\phi}(y))\}. \tag{1.16}
\]

Note that \( pl(y^*|\hat{\phi}(y)) \) above is different from the log-likelihood function computed at the MLE \((\hat{\phi}, \hat{\lambda})\), since it allows maximizing the likelihood over \( \lambda \) based on the new data \( y^* \).

Similar to the derivation of the classic AIC, assuming the asymptotic quadratic expansion like in (1.12), Xu et al. (2009) showed that an approximately unbiased estimator of pAI is given by

\[
\text{pAIC} = -2pl(y|\hat{\phi}(y)) + 2p, \tag{1.17}
\]

where \( p \) is the dimension of \( \phi \). Note that the development here is general and not specific to the PHMM, and the nuisance parameter \( \lambda \) can be either finite or infinite dimensional. If it is finite, the pAIC differs from the classic AIC that counts all the parameters in the model by twice the number of parameters in \( \lambda \).

**Marginal focus**

For the PHMM, \( p \) in (1.17) is the number of parameters in \( \beta \) and \( \Sigma \). The computation of the pAIC under the PHMM, like for the likelihood ratio, only requires the likelihood evaluated at the NPMLE, and has been discussed in the previous section.

As a special case, for the classic Cox model with no random effects, \( p \) is the dimension of \( \beta \), and the profile likelihood is the partial likelihood. This gives the AIC that has been used under the Cox model, but nonetheless had not been justified until Xu et al. (2009) to the best of our knowledge.

**Conditional focus**

For the cluster level inference, under mixed-effects models Vaida and Blanchard (2005) used the conditional log-likelihood of the observed data \( y \) given the random effects \( b \), to define a conditional Akaike information (cAI),

\[
c\text{AI} = -2E_{(y,b)}E_{y^*|b}\{l(y^*|\hat{\beta}(y), \hat{b}(y))\}, \tag{1.18}
\]
here $y^*$ is independently replicated from the same conditional distribution as the original data $y$ given the same random effects $b$, and $\beta$ contains the fixed effects. Note that this is in contrast to the marginal focus above, where new random effects $b^*$ should be sampled, which then give rise to the independently replicated data $y^*$. The estimates in (1.18) are usually assumed to be the maximum likelihood estimate of $\beta$ and the empirical Bayes estimate of $b$.

To derive the conditional AIC (cAIC), which should be an (approximately) unbiased estimate of the cAIC, one basically assumes that it is of the form

$$cAIC = -2l(y|\hat{\beta}(y), \hat{b}(y)) + 2\rho,$$

(1.19)

where $\rho$ is referred to as the bias correction term. This is because the ‘apparent’ estimate of cAIC, $-2l\{y | \hat{\beta}(y), \hat{b}(y)\}$, is known to be biased (Efron, 1986), since the estimates $\hat{\beta}(y)$ and $\hat{b}(y)$ are optimized based on the original data $y$. To find the bias correction term $\rho$, under normal linear mixed models the expectations can be computed exactly. But beyond that Donohue et al. (2011) argued that asymptotic derivation is needed, and the approach roughly follows Linhart and Zucchini (1986), albeit with parameter space that grows with the sample size (number of clusters), at least nominally.

To apply the approach of Linhart and Zucchini (1986), instead of the conditional likelihood Donohue et al. (2011) worked with the joint likelihood of $y$ and $b$:

$$l_J(y, b | \beta) = l(y | \beta, b) + \log p(b),$$

(1.20)

where $p(\cdot)$ is the density function of $b$. The joint likelihood can be seen as a penalized likelihood with $\log p(b)$ as the penalty, and is the objective function considered in Linhart and Zucchini (1986). Since $b$ is conditioned upon in the inner expectation of (1.18), it is seen that the $\rho$ is also the bias correction term for using $-2l_J(y, \hat{b} | \hat{\beta})$ to estimate $-2E_{(y,b)}E_{y^*|b}\{l_J(y^*, \hat{b}(y) | \hat{\beta}(y))\}$.

Under the PHMM the conditional log-likelihood with the baseline hazard ‘profiled’ out is the conditional partial log-likelihood

$$l(y | \beta, b) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \log \frac{\exp(\eta_{ij})}{\sum_{i'j'} \exp(\eta_{i'j'}) I(Y_{i'j'} \geq Y_{ij})}.$$

(1.21)

Let $U = (X, Z)$, where $X$ and $Z$ are the design matrices for the fixed and the random effects, and let $A = \text{diag}(0, D^{-1})$ partitioned in the same manner as $U$, where $D$ is the covariance matrix of $b$. Donohue et al. (2011) showed that

$$\rho = \text{tr}\{U^T W^* U (U^T W^* U + A)^{-1}\},$$

(1.22)

where $W^*$ was initially given in Ha and Lee (2003). The above has the same appearance as the effective degrees of freedom under the linear and generalized linear mixed
models (Hodges and Sargent, 2001; Lu et al., 2007); and this is consistent with the cAIC under the linear and generalized linear mixed models, the latter also derived in Donohue et al. (2011): cAIC is twice the negative conditional log-likelihood plus twice the effective degrees of freedom. It makes sense since the conditional focus treats the unknown but realized random effects as parameters to be estimated.

The derivation together with the conditions are given in Donohue et al. (2011). The conditions worth mentioning are that: the number of clusters $m \rightarrow \infty$, the cluster size (assuming the same size for all clusters to be simple) $n \rightarrow \infty$, and $n/m \rightarrow \infty$. The last condition is needed so that the optimizer of the joint likelihood is asymptotically equivalent to the MLE, which is typically used in practice to compute the AIC (Vonesh, 1996). It is also assume that: $\max_{i=1,...,m} \|\hat{b}_i - b_{0i}\|^2 \rightarrow 0$ almost surely as $m, n \rightarrow \infty$; $\sqrt{n}(\hat{b}_i - b_{0i}) \rightarrow N(0, v_{bi})$ in distribution for some positive definite matrix $v_{bi}$, uniformly over $i$; and $(\hat{b}_i - b_{0i})'\Delta''_{b_ib_i}(\hat{b}_i - b_{0i}) = O_p(1/m)$ as $m \rightarrow \infty$, where $\Delta''_{b_ib_i}$ is the matrix block from the Hessian matrix of $-l''_{ij}(y,b|\beta)$. The last condition was inadvertently left out in Donohue et al. (2011), plus some minor inaccuracies; a more accurate set of conditions can be found in Overholser (2013).

Example

Further continue with the E1582 lung cancer data. For the marginal focus, using the likelihood values from the previous section, it is immediate that the pAIC for the three models $d = 0, 1$ and $2$ in Table 1.2. Using the minimum pAIC $d = 2$ is chosen.

To illustrate the conditional focus, Figure 1.1 shows the estimated treatment effect from each institution. These are the empirical Bayes estimate $\hat{b}_i$'s from the model with $d = 1$. It is seen that institutions 16 and 18 have the largest treatment effects in magnitude, almost double the fixed (average) treatment effect of -0.25 (logHR). On the other hand, institutions 19, 28 and 29 how almost no treatment effects. Further investigations into how the chemo procedures were carried out at each of these institution might reveal how to improve the benefit of the alternating HEM-CAV regimens.

When the focus is on the cluster level inference as in Figure 1.1, Donohue et al. (2011) computed the cAIC under the three models $d = 0, 1$ and $2$ are given in Table 1.2. The cAIC when $d = 0$ is just the AIC under the classic Cox model using the partial likelihood, as discussed earlier. It is seen that from the conditional focus, the model with $d = 2$ is still preferred.

We should mentioned that Bayesian model selection methods have been developed in the literature under the frailty model (i.e. random intercept and random effects on binary covariates only) and under the PHMM, and have been applied to
Figure 1.1: Estimated treatment effect for each institution of the lung cancer data.

Table 1.2: Log-likelihood and AIC’s for E1582 lung cancer data

<table>
<thead>
<tr>
<th></th>
<th>$d = 0$</th>
<th>$d = 1$</th>
<th>$d = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-2 \log L$</td>
<td>7232.80</td>
<td>7228.80</td>
<td>7222.55</td>
</tr>
<tr>
<td>pAIC</td>
<td>7242.80</td>
<td>7240.80</td>
<td>7236.55</td>
</tr>
<tr>
<td>cAIC</td>
<td>6107</td>
<td>6088</td>
<td>6071</td>
</tr>
</tbody>
</table>
the E1582 lung cancer data. The Bayesian methods tend to choose only the stronger bone metastases random effect, while leaving out the random treatment effect (Dunson and Chen, 2004; Cai, 2010; Lee et al., 2012). Simulation studies in Lee et al. (2012) seem to indicate that this is due to the fact that both random effects are relatively weak. When both the variances of both the random effects are increased, the stochastic search variable selection (SSVS) method of Lee et al. (2012) ends up selecting both random effects.

1.8 PHMM: model departure and diagnostics

As illustrated in ?? one of the applications of PHMM is genetic epidemiology, an area where verifying the model assumptions and understanding the interpretations of the parameter estimates under possible model departures is of particular importance. For the standard Cox model, there has been a large body of work as summarized in ??, and in particular it is well-known that non-proportionality can lead to bias in the estimated regression effect.


In the context of PHMM, we consider extending model (1.1) to non-proportional hazards

\[
\lambda_{ij}(t) = \lambda_0(t) \exp\{\beta(t)Z_{ij} + b_i'W_{ij}\},
\]

(1.23)

where the fixed regression effect \( \beta \) is allowed to vary with time \( t \) (Xu and Gamst, 2007).

**Impact of nonproportionality on fixed effects**

Similar to Section ??, let

\[
S^{(r)}(\beta; t, \theta) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} Y_{ij}(t) \exp(\beta'Z_{ij})Z_{ij}^{\otimes r}E_\theta(e^{b_i'W_{ij}}|y_i),
\]

(1.24)

for \( r = 0, 1, 2 \), and \( a^{\otimes 0} = 1 \), \( a^{\otimes 1} = a \) and \( a^{\otimes 2} = aa' \) for a vector \( a \). Consider the NPMLE \( \hat{\theta} = (\hat{\beta}, \hat{\Sigma}, \hat{\lambda}_0) \), which is estimated using the Monte-Carlo EM algorithm as described in Section ?? . It can be shown that at the convergence of the EM
algorithm, $\hat{\beta}$ satisfies

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \int_{0}^{\infty} \left\{ Z_{ij} - \frac{S^{(1)}(\beta; t, \theta)}{S^{(0)}(\beta; t, \theta)} \right\} dN_{ij}(t) = 0. \quad (1.25)$$

This can be seen directly from the equations that are solved at the $M$-steps; it can also be derived as the score equation via direct differentiation of the log likelihood using one-dimensional submodels (Xu and Gamst, 2007).

Denote $s^{(r)}(\beta; t, \theta) = E\{S^{(r)}(\beta; t, \theta)\}, r = 0, 1, 2,$ and $v(\beta; t, \theta) = s^{(2)}(\beta; t, \theta)/s^{(0)}(\beta; t, \theta) - \{s^{(1)}(\beta; t, \theta)/s^{(0)}(\beta; t, \theta)\}$ is the derivative of $s^{(1)}/s^{(0)}$ with respect to the first argument $\beta$. We can say that the expectations are with respect to the true underlying distribution of the random variables $(T, C, Z)$; note that the random effects are an ‘intermediate’ quantity in the dependence between $T$ and $Z$, and have been ‘integrated out’ in the definition (1.24) of $S^{(r)}(\beta; t, \theta)$. The following theorem gives the population parameter that $\hat{\beta}$ converges to under model (1.23).

**Theorem 1.1** Under regularity conditions and assuming that $\int v(\beta; t, \theta)s^{(0)}(\beta(t); t, \theta)\lambda_0(t)dt$ is positive definite, as $n \to \infty$ the NPMLE $\hat{\beta}$ converges in probability to $\beta^*$, which is the unique zero of the following equation:

$$\int_{0}^{\infty} \left\{ \frac{s^{(1)}(\beta(t); t, \theta(t))}{s^{(0)}(\beta(t); t, \theta(t))} - \frac{s^{(1)}(\beta(t), \theta)}{s^{(0)}(\beta(t), \theta)} \right\} s^{(0)}(\beta(t); t, \theta(t))\lambda_0(t)dt = 0, \quad (1.26)$$

where $\theta(t) = (\beta(t), \Sigma, \lambda_0)$.

The asymptotic normality of $\hat{\beta}$ under model (1.23) was established in Dupuy (2009).

To better characterize $\beta^*$ notice the two terms inside $\{\}$ in (1.26) is the same function evaluated at $\beta(t)$ and $\beta$, respectively. With a first-order Taylor expansion we have

$$\int_{0}^{\infty} \{\beta(t) - \beta^*\} v(\tilde{\beta}(t); t, \tilde{\theta}(t))s^{(0)}(\beta(t); t, \theta(t))\lambda_0(t)dt \approx 0, \quad (1.27)$$

where $\tilde{\beta}(t)$ is between $\beta(t)$ and $\beta^*$, and $\tilde{\theta}(t) = (\tilde{\beta}(t), \Sigma, \lambda_0)$. Solving (1.27) for $\beta^*$ we have

$$\beta^* \approx \frac{\int \beta(t)v(\tilde{\beta}(t); t, \tilde{\theta}(t))s^{(0)}(\beta(t); t, \theta(t))\lambda_0(t)dt}{\int v(\tilde{\beta}(t); t, \tilde{\theta}(t))s^{(0)}(\beta(t); t, \theta(t))\lambda_0(t)dt}. \quad (1.28)$$

This way $\beta^*$ is a weighted average of $\beta(t)$ over time. Meanwhile from the $dN_{ij}(t)$ term in (1.25) we see that when there is censoring, the censored observations lose their contributions to the estimating equation. In other words, the equation gives
insufficient weights to the (later) censored observations. This results in an average $\beta$ value that is biased towards the earlier values of $\beta(t)$, as compared to the uncensored case. In the simulations below we will see that this is precisely the case. Notice that this is also consistent with the fixed-effects-only case discussed earlier in Chapter ??.

**Impact of nonproportionality on variance components**

The approach above for $\hat{\beta}$ does not provide direct insight for the behavior of $\hat{\Sigma}$ under model (1.23). Also, the effect of non-proportional hazards on $\hat{\Sigma}$ is much more complex, as illustrated by numerical studies.

Xu and Gamst (2007) considered simple scenarios as follows. The more complex situation may be seen as mixtures of the simpler scenarios. They considered the special case of a single covariate with both non-proportional and random effects. Suppose that the underlying model is

$$
\lambda_{ij}(t) = \lambda_0(t)e^{\beta(t)+b_i}Z_{ij},
$$

(1.29)

while we fit

$$
\lambda_{ij}(t) = \lambda_0^*(t)e^{(\beta^*+b_i^*)}Z_{ij}.
$$

(1.30)

Here the random effects $b_i$'s are also seen as unknown quantities to be estimated, the so-called Empirical Bayes point of view. Results of Section ?? imply that in the absence of censoring the 'least-false' parameter value

$$
\beta^* + b_i^* \approx \int \{\beta(t) + b_i\}dF_i(t) = \int \beta(t)dF_i(t) + b_i,
$$

(1.31)

where $F_i(t)$ is the marginal distribution function of the failure times in cluster $i$. Without loss of generality we may assume that the covariate values are non-negative, since in practice these values are usually bounded. Then a larger $b_i$ implies higher relative risks and shorter failure times in cluster $i$, i.e. $F_i(\cdot)$ puts more weight on earlier times. If $\beta(t)$ is decreasing, this shows that $b_i$ and $\int \beta(t)dF_i(t)$ are positively correlated, therefore $\text{Var}(b_i^*) = \text{Var}(\beta^* + b_i) > \text{Var}(b_i)$. When allowing censoring in general, $F_i(\cdot)$ is replaced by the intensity of the counting process, which also puts more weight on the earlier values of $\beta(t)$ under larger $b_i$ and higher relative risks. This shows that there will be overestimation of the variance component if $\beta(t)$ is decreasing. On the other hand, if $\beta(t)$ is increasing, $b_i$ and $\int \beta(t)dF_i(t)$ will be negatively correlated, leading to underestimation of the variance component. These analytical considerations are confirmed by the simulation results below.
Numerical results

Here we reproduce Table 1 from Xu and Gamst (2007). The results of Table 1.3 were generated with 40 clusters and 10 observations each cluster. In all three scenarios with either random intercept or random slope, increasing or decreasing $\beta(t)$, we see that under PHMM $\hat{\beta}$ estimates an average $\beta(t)$ value that lies between $\beta_0$ and $\beta_1$. With increasing amount of censoring, this average value is shifted more towards the earlier value $\beta_0$ of $\beta(t)$. This confirms our previous conclusions based on the analytic expressions (1.25) and (1.26).

For the effect on the variance component, in the first scenario of Table 1.3, we had a random intercept in the log hazard, i.e. $w = 1$. The estimated variance $\hat{\sigma}^2$ of the random effects, in this case, does not appear to be affected by the non-proportionality, in particular when compared to the empirical variance $v(b)$ of the simulated random effects. For the rest of the table we had a random effect on $w = Z$. In the second scenario with decreasing $\beta(t)$ the estimated variance of the random slope $\hat{\sigma}^2$ overestimates the true variance, while in the third scenario with increasing $\beta(t)$ it underestimates as we have derived analytically. In summary, the non-proportionality does not affect the variance of the random intercept, which is the random effect on the log baseline hazard, but does affect the variance of the

Table 1.3: Effect of non-proportional hazards on the parameter estimates (single covariate)

<table>
<thead>
<tr>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>Random</th>
<th>% censored</th>
<th>$\beta$</th>
<th>$\hat{\sigma}^2$</th>
<th>$v(b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0%</td>
<td>0.239 (0.107)</td>
<td>0.939 (0.277)</td>
<td>0.965 (0.234)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>33%</td>
<td>0.318 (0.150)</td>
<td>0.982 (0.303)</td>
<td>1.020 (0.235)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>0.403 (0.158)</td>
<td>0.943 (0.310)</td>
<td>0.958 (0.204)</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0%</td>
<td>0.231 (0.209)</td>
<td>1.204 (0.390)</td>
<td>0.965 (0.234)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>33%</td>
<td>0.267 (0.227)</td>
<td>1.232 (0.407)</td>
<td>1.020 (0.235)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>0.315 (0.229)</td>
<td>1.167 (0.432)</td>
<td>0.958 (0.204)</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0%</td>
<td>0.876 (0.184)</td>
<td>0.771 (0.262)</td>
<td>0.965 (0.234)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>33%</td>
<td>0.822 (0.193)</td>
<td>0.783 (0.300)</td>
<td>1.020 (0.235)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>0.781 (0.227)</td>
<td>0.740 (0.289)</td>
<td>0.958 (0.204)</td>
<td></td>
</tr>
</tbody>
</table>

$\lambda_{ij}(t) = \exp\{\beta(t)Z_{ij} + b_iw_{ij}\}, i = 1, ..., 40, j = 1, ..., 10. \beta(t) = \beta_0$ when $t < 0.1$ and $\beta_1$ otherwise, $b \sim N(0, 1). \ Z$ is binary 0,1 with equal probabilities. Uniform $(0, \tau)$ censoring. $v(b)$ is average sample variance of $b$ from the simulations. In the $(\cdot)$ are standard errors from 100 simulations.
random slope on the same covariate.

Additional numerical results in Xu and Gamst (2007) give further indications for more complex cases. In general, it appears that for a covariate with time-varying effect $\beta(t)$, its estimated $\hat{\beta}$ is an average effect that can be affected by censoring; the variance of its random effect is under- or overestimated depending on whether $\beta(t)$ increases or decreases over time. The effect of non-proportionality of one covariate on other covariates are more mixed. For independent covariates, it appears that the non-proportionality of one covariate does not affect the estimated fixed effects of other covariates; this has also been observed in Strandberg et al. (2013) when there are no random effects in both the data generating model and the fitted model. The variance components corresponding to other covariates, however, could be either affected or unaffected.

Checking the proportional hazards assumption

Xu and Gamst (2007) developed a method using cumulative residuals similar to those in (Fleming and Harrington, 1991, page 176), in order to checking the proportional hazards assumption. Given the estimates $\hat{\beta}$ and $\hat{b}_i$, define the standardized covariate residual

$$r_{ij} = \hat{V}(X_{ij})^{-1/2} \left\{ Z_{ij} - \frac{\hat{S}^{(1)}(X_{ij})}{\hat{S}^{(0)}(X_{ij})} \right\}, \quad (1.32)$$

where

$$\hat{S}^{(r)}(t) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} Y_{ij}(t) \exp(\hat{\beta}' Z_{ij}) Z_{ij}^{\otimes r} e^{\hat{b}_i W_{ij}} \quad (1.33)$$

for $r = 0, 1, 2$, and $\hat{V}(t) = \hat{S}^{(2)}(t)/\hat{S}^{(0)}(t) - \{\hat{S}^{(1)}(t)/\hat{S}^{(0)}(t)\}^{\otimes 2}$. If the PHMM holds, the components of $r_{ij}$ should have approximately mean zero and variance one. Here we implicitly assume that the estimates are ‘reasonably good’. Jiang (1998) discussed the empirical Bayes estimates of random effects under linear mixed models, and the residuals formed using these. It showed that they work well when the cluster sizes are reasonably large, which makes sense intuitively since there would be enough sample to estimate $b_i$ from cluster $i$. In our case we found the method below to work well even when data contain some small clusters such as with only one observation.

As reviewed in (O’Quigley and Xu, 1998, page 1736-7), residuals like the above may not be sensitive to departures from the proportional hazards assumption. On the other hand, we may form cumulative residuals over time, which seems more sensitive for model checking. Let $r_{(l)}$ be the $l$-th ordered residual corresponding to
1-th failure time in the data, assuming no ties for simplicity. Denote the cumulative residuals \( C_k = \sum_{l=1}^{k} r_l \), \( k = 1, \ldots, K \) where \( K \) is the total number of failures. Then the components of \( C_k \), \( k = 1, \ldots, K \), are approximated by the Brownian Motion under PHMM. We may further standardize by letting \( C(k/K) = C_k / \sqrt{K} \), and the whole process \( C(s) \) \( (0 \leq s \leq 1) \) is obtained by linear interpolation between \( s = (k - 1)/K \) and \( s = k/K \). The components of \( B(s) = C(s) - sC(1) \) are then approximated by the Brownian Bridge on \([0, 1]\). It is known that for Brownian Bridge

\[
P \left( \max_{0 \leq s \leq 1} |B_p(s)| \geq a \right) \approx 2 \exp(-2a^2), \quad p = 1, \ldots, P,
\]

where \( P \) is the dimension of \( Z \). In particular, if we plot the Brownian Bridge computed as in the above from a dataset, and if its maximum absolute value exceeds 1.36 (or 1.63), then we may reject the proportional hazards assumption under PHMM at 0.05 (or 0.01) significance level for this data set.

**Example**

We continue with the E1582 lung cancer data set. Here we consider model (1.1) with fixed effects for all five covariates including treatment, and a random effect for treatment.

In Figure 1.2 top row are the standardized covariate residuals for each covariate. Notice that all five covariates are binary, leading to residual plots that are particularly not informative sometimes. A closer look shows that only performance status has more than 5% of the residuals outside of \( \pm 1.96 \). The second row of the figure shows components of the Brownian Bridge processes as defined above. The maximum absolute values of the Brownian Bridges are: treatment 1.10, bone metastases 1.23, liver metastases 1.70, performance status 2.04, and weight loss 0.86. This indicates a significant departure from the proportional hazards assumption for the fixed effects of liver metastases and performance status. The third row shows the estimated \( \beta(t) \)'s using the penalized splines approach that will be described in the next section. We note that the estimated \( \beta(t) \) for liver metastases and performance status exhibits substantial decreasing and increasing trend over time, while the effects of the other three covariates are relatively stable.

1.9 Partially proportional hazards mixed-effects model

Here we consider estimation under model (1.23). We keep the same normal distribution assumption on the random effects as before. The time-varying fixed effects \( \beta(t) \) can be approximated, among other choices, splines. One advantage of using the spline basis, is the algebraic equivalence between fitting a spline function with
1.9. **PARTIALLY PROPORTIONAL HAZARDS MIXED-EFFECTS MODEL**

Figure 1.2: Standardized covariate residuals, Brownian bridges, and estimated $\beta(t)$ for the lung cancer data.
penalty (for given penalty or smoothing parameter) and fitting a random effects model given the variance components: the best linear unbiased predictor (BLUP) of the random effects is the same as the estimated spline coefficients (Speed, 1991; Brumback and Rice, 1998; Ruppert et al., 2003). In this way the time-varying fixed effects and the random effects can be treated in the same framework.

More specifically, Kauermann et al. (2008) considered the penalized splines (p-splines) for \( \beta(t) \), as well as for the baseline hazard function \( \log \lambda_0(t) \). With a slight change of notation, we rewrite (1.23) as

\[
\lambda ij(t) = \exp \{ \beta(t)'Z_{ij} + b_iW_{ij} \},
\]

(1.34)

where \( Z \) now has a first component of ‘1’, and \( \beta_1(t) = \log \lambda_0(t) \). Assume that \( \beta_k(t) \approx \beta_{0k} + a_k B(t) \) for \( k = 1, ..., p + 1 \), where \( B(t) \) is a high-dimensional spline basis; a constant component of \( \beta(t) \) over time can be treated as a special case. Here for simplicity we use the same basis for all components of \( \beta(t) \). An example of \( B(t) \) is the truncated polynomials basis: \( B_k(t) = \{ (t - t_0) + (t - t_1) + \ldots, (t - t_q) \} \), where \( t_0 < t_1 < \ldots < t_q \) are fixed knots covering the range of the observed failure times.

Conditional on the random effects \( b \), the log-likelihood is

\[
l_c(\beta_0, a, b) = \sum_{i=1}^m \sum_{j=1}^{n_{ij}} \left[ \delta_{ij} \{ \beta_{0i}'Z_{ij} + Z_{ij}'B(t_{ij})a + b_iW_{ij} \} - \int_0^{X_{ij}} \exp \{ \beta_{0i}'Z_{ij} + Z_{ij}'B(t)a + b_iW_{ij} \} dt \right].
\]

(1.35)

where \( B(t) \) is a block diagonal matrix with \( p + 1 \) blocks of \( B(t) \). Since the number of knots, \( q \), are generously chosen for p-splines, a penalty term needs to be added to the log-likelihood above. In Kauermann et al. (2008) \( q = K/2 \) is used where \( K \) is the total number of events. As mentioned earlier the penalization is equivalent to treating the spline coefficients as equivalent to an \( a \) priori distribution:

\[
a \sim N(0, \Sigma_a),
\]

(1.36)

where \( \Sigma_a = \text{diag}(\sigma_k^2 D)_{k=1}^{p+1} \) for some \( D \). \( D = I \) identity matrix corresponds to a penalty of the form \( \sum_{i=1}^q a_i^2 \), and \( \sigma_k^2 \) corresponds to the penalty or smoothing parameter for the \( k \)th component of \( \beta(t) \).

The second component in (1.35) is the integrated hazard. By approximating \( \log \lambda_0(t) \) using splines, this integral does not become a sum like under the PHMM. Kauermann et al. (2008) approximated the integral by the rectangle method using the observed failure time points as a partition of the time axis. The resulting likelihood was then recognized as one from a pseudo Poisson data set, in a way similar to the equivalence between the Cox model and a Poisson likelihood (Whitehead, 1980). Donohue et al. (2011) pointed out the computational inefficiency of using
1.9. **PARTIALLY PROPORTIONAL HAZARDS MIXED-EFFECTS MODEL**

pseudo Poisson data to fit the PHMM. As an alternative to the Kauermann *et al.* (2008) approach, it might be advantageous to still consider the NPMLE for \( \Lambda_0 \), so that the likelihood remains one under the PHMM, after recognizing that the spline coefficients in \( a \) can be treated as random effects via (1.36). On the other hand, since \( \sigma^2_k \) corresponds to the penalty or smoothing parameter, automated PHMM package should not be used to ‘estimate’ the \( \sigma^2_k \)’s in our opinion. [footnote: there has some advocacy in the literature under this kind of setting to ‘estimate’ the smoothing parameters using the restricted maximum likelihood (REML), and some mathematical arguments are given, but no sampling theory has been established for the results estimates to the best of our knowledge.]

Although \( a \) and \( b \) can both be treated formally as random effects, their asymptotic scenarios can be different. Typically for clustered data, we require that \( m \to \infty \) in theory, and in practice often the number of clusters is large when we used mixed-effects model. On the other hand, the cluster sizes \( n_i \), may or may not be large. This implies that the sample size for estimating the spline coefficients \( a \) is typically large, but that is not necessarily the case for estimating the original random effects in \( b \). Based on these considerations, Kauermann *et al.* (2008) proposed to integrate out \( a \) from the conditional likelihood above using Laplace approximation, while integrating out \( b \) using the EM algorithm. The resulting algorithm, which they called stacked Laplace-EM, is then computationally efficient and accurate.

Finally for selection of the smoothing parameters, criteria such as AIC can be considered. The usual AIC used for smoothing, is the same as the cAIC when the spline coefficients are treated as random effects. Overholser (2013) explored the use of AIC versus BIC in linear models with both random and smooth effects. It is certainly an area worth further investigation.

**Example**

Model (1.23) was fitted to the E1582 lung cancer data set using the method of Kauermann *et al.* (2008) described above. The resulting \( \beta(t) \) for the five covariates are shown in Figure 1.2, together with 95% pointwise confidence intervals obtained using sandwich estimate for variance of the parameters under the penalized log-likelihood.
Bibliography


