Patrick Mair and Marcus Hudec

Session Clustering Using Mixtures of Proportional Hazards Models

Working Paper

Session Clustering Using Mixtures of Proportional Hazards Models

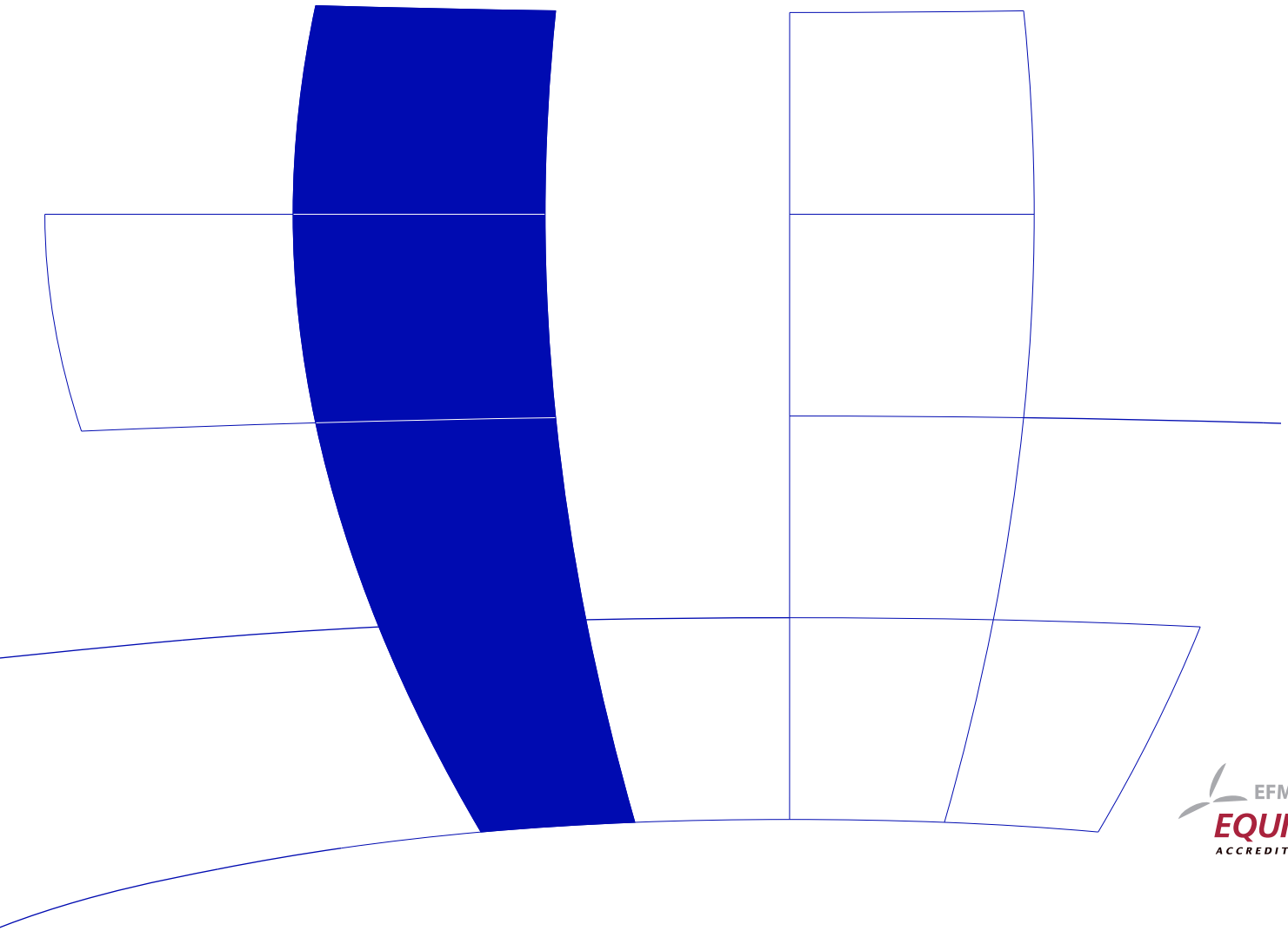
Patrick Mair, Markus Hudec

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 63
March 2008

http://statmath.wu-wien.ac.at/



Session Clustering Using Mixtures of Proportional Hazards Models

Patrick Mair
Wirtschaftsuniversität Wien

Marcus Hudec
University of Vienna

Abstract

Emanating from classical Weibull mixture models we propose a framework for clustering survival data with various proportionality restrictions imposed. By introducing mixtures of Weibull proportional hazards models on a multivariate data set a parametric cluster approach based on the EM-algorithm is carried out. The problem of non-response in the data is considered. The application example is a real life data set stemming from the analysis of a world-wide operating eCommerce application. Sessions are clustered due to the dwell times a user spends on certain page-areas. The solution allows for the interpretation of the navigation behavior in terms of survival and hazard functions. A software implementation by means of an R package is provided.

Keywords: proportional hazards models, Weibull mixture models, EM-algorithm, incomplete data, Web usage mining.

1. Introduction

In this paper we extend the idea of probabilistic clustering to observed survival times or, as in our example, dwell times. Within a web mining context we postulate that there exist a number of different segments of users generating different dwell time patterns over the sessions. Within each segment the user behavior with respect to the dwell times is similar in some sense; between these segments the user behavior is different. Since the segments themselves are unknown, our approach leads us to the problem of unobserved heterogeneity within the context of survival data (Heckman and Singer 1984; Honoré 1990) and it is solved by means of a parametric clustering approach.

To achieve a parametric clustering some assumptions concerning the distribution have to be established. Our core assumption is that dwell times are Weibull distributed with different parameters over the clusters (i.e., a Weibull mixture model). Methodological and application issues are elaborated in the next sections. In addition, we propose a more parsimonious modeling strategy by using mixture Weibull proportional hazards models (WPHM). To estimate the parameters, the EM-algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Krishnan 1997; Ramon, Albert, and Baxter 1995) is used and since not every page is visited by every user, we have to take into account the problem of incomplete data.

From the application perspective we focus on the segmentation of users according to their dwell times on various areas of a web-site. This means that we aggregate the often complex and spacious topology of the site into a set of non-overlapping areas. A session is then characterized by a vector of dwell times a user spends when visiting the according area of the site.

Recent research in web usage mining has focused on various methods of modeling user navigation history on web sites. Most of these approaches deal with the traversal order of pages within a session. The methodology applied covers a wide spectrum: hidden Markov chain models, sequence alignment methods, hypertext probabilistic grammars, semi-structured temporal graphs, and sequential association analysis. Notable works are provided by Smyth (1999), Cadez, Heckerman, Meek, Smyth, and White (2001), and Ypma and Heskes (2002). The main idea behind the men-

tioned papers is the application of model-based clustering procedures which are primarily defined by mixture models. Each single mixture component corresponds to a cluster. Within these components the modeling idea is that of a Markov chain, i.e., each session is regarded as a finite state Markov model. Clusters are formed on the base of the navigation pattern (clickstream) and the resulting transition probabilities.

However, these approaches do not take into account the dwell times. The potential of approaches focusing on dwell time analyses is well-known in methodological marketing literature. For instance, Montgomery, Li, Srinivasan, and Liechty (2004) propose a dynamic multinomial probit model of navigation patterns which leads to a remarkable increase of conversion rates. Park and Fader (2004) developed multivariate exponential gamma models which enhance cross-site customer acquisition.

In this paper, first, the analogy of "dwell times" in web usage mining and "survival times" in medical statistics is taken into consideration. This view opens a large framework of well-known survival models for the analysis of web usage patterns. In a second step methodological issues pertaining to multivariate mixtures of Weibull proportional hazard models with incomplete data are developed. Dealing with missing values is an important aspect for our application since usually a user does not visit all possible pages on a host. First elaborations can be found in Mair and Hudec (2008). All practical computations are performed with the `mixPHM` package (Mair and Hudec 2007) in R (R Development Core Team 2008).

2. Webshop Data for Session Dwell Times

The whole methodology as described in Section 3 is driven by a project in cooperation with a world-wide operating Austrian eCommerce company. The main goal of this running project is to develop successful web mining strategies in the context of experimental research and to implement the analysis tools into the business process to improve the efficiency of their shop in the long run.

The data collection and preparation is carried out in an automated way. Basically, when visiting the company host, each subject leaves a "trace" by means of a server protocol. In web mining applications this type of data is commonly referred to as "clickstream". Hence, the source of the data are log-files (using pixel-log methodology) provided by the company's webserver. These log-files include various fields such as date, time, client IP, server IP, cookie information etc. in a rather unstructured manner. At the end each log-file is a string which has to be parsed in order to analyze the data statistically. This parsing process is carried out by means of a customized ETL-tool (Extract, Transform, Load). Note that the server provides one log-file for each day. Each page impression (PI) accomplished by a visitor of the shop produces one line in the log-file.

Subsequent to the ETL process each pixel-log is structured in a file with over 100 fields. Having this raw file on a PI-level integrated into a database, the following crucial steps are executed in order to have a flat file appropriate for dwell time analysis.

The first step is to coarsen the topology of the web-site which has a granularity much too fine (e.g., a single web page for each offered product), as to find reasonable navigation paths. Hence, the single pages are categorized into page-types (i.e., site-areas) by means of content-driven considerations. For instance, pages of products of a certain type are merged into one site-area, all pages during the checkout process are combined, etc.).

The second relevant information for our purposes is the time stamp of the PI. Based on the time that a particular page A is opened until the time the next page B is called, the dwell time for page A can be computed straightforwardly. At the end of this step we have an assignment to a particular site-area and the corresponding dwell time for each PI. Anticipatory to the subsequent computation of the likelihood it is important to point out that in our example the dwell times between different page areas do not show any noticeable correlation structure.

In most web mining applications the researcher is not interested in doing analyses on a PI-level but rather on a session/user level. By using cookies it is easy to identify which PI belongs to which session. The user identification based on multiple sessions is somewhat more tricky but in

practice it is accomplished again by cookies or by IP (if it is static). We focus our analyses on the user level but it is straightforward to apply our algorithm to a session level; it is just a matter of data aggregation/preparation.

Once having the data at a session level the final aggregation step is carried out. Obviously, a particular page-type can be visited more times during a session. In our approach we want to cluster the sessions due to their joint dwell times on the site-areas and thus, corresponding multiple dwell times on the same page-type are added. Finally, we have a flat file of $i = 1, \ldots, N$ unique sessions in the rows and $p = 1, \ldots, P$ site-areas in the columns (see Section 4). Since not each page-type will be visited by each user, our approach must be able to handle a large amount of informative missing values. These missings provide important information for achieving the final cluster solution, namely that a certain page-type was not visited within the corresponding session.

3. Model Specification and Estimation

3.1. Weibull Mixtures and Proportional Hazard Models

Since survival analysis focuses on duration times until some event occurs (e.g., the death of a patient in medical applications) it seems straightforward to apply these concepts to the analysis of dwell times in web usage mining applications.

With regard to dwell time distributions we assume that they follow a Weibull distribution with density function $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})$, where λ is the scale parameter and γ the shape parameter. To model the heterogeneity of the observed population, we assume K latent segments of sessions/users. Since the Weibull assumption holds within all segments, different segments exhibit different parameter values. This leads to the underlying idea of a *Weibull mixture model*. For each page category p the resulting mixture density is of the following form:

$$f(t_p) = \sum_{k=1}^{K} \pi_k f(t_p; \lambda_{pk}, \gamma_{pk}) = \sum_{k=1}^{K} \pi_k \lambda_{pk}\gamma_{pk} t_p^{\gamma_{pk}-1}\exp(-\lambda_{pk} t_p^{\gamma_{pk}}) \tag{1}$$

where t_p represents the dwell time on page category p with mixing proportions $\pi_k > 0$ which corresponds to the relative size of each segment k such that $\sum_{k=1}^{K}\pi_k = 1$.

Assuming that the dwell times over various page areas are independent, the joint likelihood expression can be determined straightforwardly. The parameters are estimated with the EM-algorithm and corresponding identifiability and convergence issues can be found in Ishwaran (1996) and Jewell (1982).

To reduce the number of parameters involved we impose restrictions on the hazard rates of different mixture components and pages, respectively. An common way of doing this is offered by the concept of proportional hazards models (PHM) as for instance given in Kalbfleisch and Prentice (1980), Cox and Oakes (1984), Collett (2003):

$$h(t; \mathbf{z}) = h_0(t)\exp(\mathbf{z}\boldsymbol{\beta}). \tag{2}$$

This model assumes that the underlying hazard rate is a function of the baseline hazard $h_0(t)$ and of covariates $\mathbf{z}$ with corresponding regression parameters $\boldsymbol{\beta}$. Based on the distribution assumption $h_0(t)$ is specified. For the Weibull distribution the PHM above becomes

$$h(t; \boldsymbol{z}) = \lambda\gamma t^{\gamma-1}\exp(\boldsymbol{z}\boldsymbol{\beta}) \tag{3}$$

which is called WPHM. If the Weibull assumption may not be appropriate, on the one hand less parameterized distribution such as exponential and Rayleigh can be modeled. On the other hand less restrictive models such as semi-parametric Cox regression (Cox 1972) can be used where $h_0(t)$ is nonparametric. In WPHM the model parameters λ, γ, and $\boldsymbol{\beta}$ can be estimated jointly by maximum likelihood.

3.2. Parsimonious Modeling Strategies

Within the context of our data example we propose five different models with respect to different proportionality restrictions in the hazard rates. Hence, by imposing such restrictions the parameters are reduced with respect to the Weibull mixture model. In order to have a common WPHM framework, the Weibull mixture model can be stated as WPHM: The hazard of session s_i belonging to component k on page category p is

$$h(t_{i,p}; \mathbf{1}) = \lambda_{k,p}\gamma_{k,p}t_{i,p}^{\gamma_{k,p}-1}\exp(\boldsymbol{\beta}\mathbf{1}). \tag{4}$$

The parameter matrices can be represented jointly as

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{1,1} & \dots & \lambda_{1,P} \\ \vdots & \ddots & \vdots \\ \lambda_{K,1} & \dots & \lambda_{K,P} \end{pmatrix}$$

for the scale parameters and

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{1,1} & \dots & \gamma_{1,P} \\ \vdots & \ddots & \vdots \\ \gamma_{K,1} & \dots & \gamma_{K,P} \end{pmatrix}$$

for the shape parameters. Both the scale and the shape parameters can vary freely and there is no assumption of hazard proportionality. The number of parameters is $2 \times K \times P$ and they correspond to Weibull mixture model parameters.

The first restriction we impose is proportionality of hazards across groups. This can be modeled by imposing the latent component vector $\boldsymbol{g}$ as a contrast. This vector is of length N and assigns each session to a cluster k. Hence, we model $h(t; \boldsymbol{g})$. Again, the elements of the matrix $\boldsymbol{\Lambda}$ of scale parameters can vary freely, whereas the shape parameter matrix reduces to the vector $\boldsymbol{\Gamma} = (\gamma_{1,1}, \dots, \gamma_{1,P})$. Thus, the shape parameters are constant over the components and the number of parameters is reduced to $K \times P + P$.

If we impose page contrasts in the WPHM, i.e., $h(t; \boldsymbol{p})$, as before the elements of $\boldsymbol{\Lambda}$ are not restricted at all but this time the shape parameters are constant over the pages, i.e., $\boldsymbol{\Gamma} = (\gamma_{1,1}, \dots, \gamma_{1,K})$. The total number of parameters is now $K \times P + K$.

The most restrictive model is the main-effects model $h(t; \boldsymbol{g} + \boldsymbol{p})$ where we impose proportionality restrictions on both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ such that the total number of parameters is reduced to $K + P$. For the scale parameter matrix proportionality restrictions of this model hold row-wise as well as column-wise:

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & c_2\lambda_1 & \dots & c_P\lambda_1 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_K & c_2\lambda_K & \dots & c_P\lambda_K \end{pmatrix} = \begin{pmatrix} \lambda_1 & \dots & \lambda_P \\ d_2\lambda_1 & \dots & d_2\lambda_P \\ \vdots & \ddots & \vdots \\ d_K\lambda_1 & \dots & d_K\lambda_1 \end{pmatrix}.$$

The c- and d-scalars are proportionality constants over the pages and components, respectively. The shape parameters are constant over the components and pages. Thus, $\boldsymbol{\Gamma}$ reduces to one shape parameter γ which implies that the hazard rates are proportional over components and pages.

To relax the rather restrictive assumption with respect to $\boldsymbol{\Lambda}$ we can extend the main effects model by the corresponding component-page interaction term, i.e., $h(t; \boldsymbol{g} * \boldsymbol{p})$. The elements of $\boldsymbol{\Lambda}$ can vary freely whereas $\boldsymbol{\Gamma}$ is again reduced to one parameter only, leaving us with a total number of parameters of $K \times P + 1$. With respect to the hazard rate this relaxation implies again proportional hazards over components and pages.

3.3. EM-Estimation of the Mixture Proportional Hazard Models

Since our approach is primarily intended for modeling dwell times on web pages, we have to take into account that many observations will not have a dwell time because certain pages are not

frequented by users. In other words, not many users will visit each page-area. This incomplete data problem is solved in the following way.

In mixture modeling the maximum likelihood equation typically consists of the joint density function composed of the single components with the corresponding mixture weights (McLachlan and Peel 2000). For establishing the likelihood function, i.e., taking the product over the joint probability density over all individuals (sessions), it is required that the data in the multidimensional space are complete over all variables. Otherwise, the session is excluded case-wise. For our case this would mean that nearly every session would be excluded from the analysis since very few users visited all page categories. Setting a missing failure time equal to zero would not be feasible, since the Weibull distribution is defined only for $t > 0$.

To solve this issue we introduce a "prior" probability that an element (session) s_i of component k visits page p. It is denoted by $\tau_p(s_i|k)$ and estimated by the corresponding relative frequency. Taking into account the Weibull dwell-time assumption, the resulting "posterior" probability $\phi_p(s_i|k)$ for session s_i being in component k (for each page p) is

$$\phi_p(s_i|k) = \begin{cases} f(t_p; \hat{\lambda}_{k,p}, \hat{\gamma}_{k,p})\tau_p(s_i|k) & \text{if } p \text{ was visited by } s_i \\ 1 - \tau_p(s_i|k) & \text{if } p \text{ was not visited by } s_i \end{cases} \tag{5}$$

where t_p are the observed dwell times on page p. Correspondingly, $f(t_p; \hat{\lambda}_{k,p}, \hat{\gamma}_{k,p})$ is the Weibull density with parameters $\hat{\lambda}_{k,p}$ and $\hat{\gamma}_{k,p}$ estimated using WPHM presented above. Note that prior/posterior are double-quoted since they are not priors/posteriors in the classical EM-context such as in (7). To establish the page joint likelihood, independence of the dwell times over pages is assumed and thus,

$$L(s_i|k) = \prod_{p=1}^{P} \phi_p(s_i|k). \tag{6}$$

By looking at each session s_i separately, a vector of likelihood values

$$\Psi_i = (L(s_i|k=1), L(s_i|k=2), \ldots, L(s_i|k=K))$$

results. These are the final values in the E-step: The likelihood that a certain session s_i belongs to group $k = 1, \ldots, K$.

In the subsequent M-step there are several possibilities of achieving an assignment of a session to a component. The classical EM-algorithm computes the posterior probabilities

$$\nu(s_i|k) = \frac{L(s_i|k)}{\sum_{k=1}^{K} L(s_i|k)} \tag{7}$$

that session s_i is assigned to component $k = 1, \ldots, K$. Straightforwardly the posterior computation implies that $\sum_{k=1}^{K} \nu(s_i|k) = 1$. By applying this classical EM-strategy, at the end we have a probabilistic or "soft" assignment of the sessions to the components.

Alternative ways such as proposed by Celeaux and Govaert (1992) provide a deterministic or "crisp" cluster assignment by a modified M-step: In their *classification EM* (CEM) each session is assigned deterministically to a cluster due to the maximal posterior probability. If the posteriors are quoted as vector

$$\boldsymbol{\nu}(s_i) = (\nu(s_i|k=1), \nu(s_i|k=2), \ldots, \nu(s_i|k=K)), \tag{8}$$

the assignment for each s_i is carried out due to $\sup_k(\boldsymbol{\nu}(s_i))$. Note that in the CEM the computation of the posterior matrix $\boldsymbol{\nu} = (\boldsymbol{\nu}(s_1), \boldsymbol{\nu}(s_2), \ldots, \boldsymbol{\nu}(s_N))$ can be omitted since the group assignment can be achieved by $\sup_k(\boldsymbol{\Psi}_i)$. This strategy leads to a remarkable decrease in computation time and hence, if a probabilistic cluster assignment is not necessarily needed, this strategy is recommended for large data set applications typical for web mining. Alternatively, a stochastic CEM version provides a corresponding randomized group assignment which takes into account the probability values of the posteriors.

Irrespective of the M-strategy used, the joint likelihood value can be computed by means of the following steps. Let us denote

$$\widetilde{L}(s_i) = \sup_k (\boldsymbol{\Psi}_i) \tag{9}$$

as the maximum likelihood value for s_i over the groups. Correspondingly, the joint log-likelihood ℓ becomes

$$\ell = \sum_{i=1}^{N} \log \widetilde{L}(s_i). \tag{10}$$

The EM iterations are carried out in the following way: We start with an initial group assignment vector $\boldsymbol{g}^{(0)}$ for CEM or a posterior matrix $\boldsymbol{\nu}^{(0)}$ for classical EM. Within each iteration l the parameter matrices $\widehat{\boldsymbol{\Lambda}}^{(l)}$ and $\widehat{\boldsymbol{\Gamma}}^{(l)}$ are estimated in the E-step. Based on these estimators the session-wise likelihood values $\Psi_i^{(l)}$ are computed. In the M-step the group vector $\boldsymbol{g}^{(l)}$ and $\boldsymbol{\nu}^{(l)}$, respectively, are updated which in turn act as starting values for iteration $l+1$. Finally, the log-likelihood value $\ell^{(l)}$ is calculated. The iteration stops when a convergence criterion such as $\left|\ell^{(l)} - \ell^{(l-1)}\right| \leq \epsilon$ is reached.

4. Clustering Webshop Users

Now we demonstrate a prototypical way of performing parametric session/user clustering with corresponding interpretation of various restricted models. All computations are performed with the `mixPHM` package in R. Emanating from the data file prepared according the steps described in Section 2 a sample of 10000 users is drawn. To be able to visualize the results properly, we limit our computations to a 6-cluster solution and to the following 6 site areas: search page which allows for searching products, gift finder where one can get a recommendation of gifts for special occasions, three areas of product groups, and checkout which includes all the pages from the shopping basket until the final payment. Note that due to nondisclosure agreements the product pages are made anonymous. The input data structure `webdata` ($N = 10000$, $P = 6$) with the dwell times in seconds is of the following form:

```
      search giftfinder product1 product2 product3 checkout
User1    785       2180      127     3388        2       NA
User2     94         63       51     1434      174      450
User3     NA        115      195       NA       99       NA
User4     70          8      546      621       10       NA
User5     79       1301       79        4        4       NA
...
```

In this section we focus on the interpretation of the different models in the context of web usage mining. For $K = 6$ all different models are computed and the hazard behavior will be examined. The mixture Weibull model will be denoted as M_{sep} since all parameters are estimated separately from each other. It is the most general model. M_p includes the page contrasts; the hazard rates are proportional over pages. M_g includes the group contrast; the hazards are proportional over groups. M_{p+g} is the main effects model which leads to hazard proportionality over groups and pages and M_{p*g} accounts for interactions. First, we look closer at the results of the unrestricted M_{sep}. The cluster mean dwell times for represented as profile plot in Figure 1 are the following:

```
          search giftfinder product1 product2 product3 checkout
Cluster 1  41.95     116.57    40.36   787.12    19.03     9.73
Cluster 2  89.07     162.81   404.12   809.56    50.25   662.87
Cluster 3  46.18      98.16   398.56    38.67    55.76     3.73
Cluster 4 307.65     331.93   677.49   858.63   206.03     4.89
Cluster 5   7.99      83.33   497.35  1118.92   158.33   270.02
Cluster 6  44.03     128.79   519.87   527.63   115.81   135.93
```

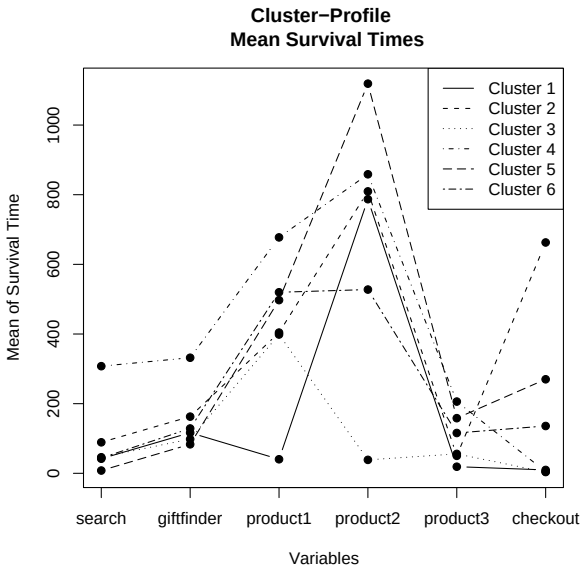


Figure 1: Cluster Profiles

Inspecting the mean dwell times we see that members of Cluster 4 have throughoutly high dwell times except for the checkout page. The highest dwell time on this page is achieved by Cluster 2 followed by Cluster 5. Cluster 1 and Cluster 3 have considerably low dwell times depending on the page-type. For an external validation the covariate "buyer" (yes/no) is taken into account. A corresponding cross-classification with the deterministic cluster solution of M_{sep} (and all subsequent models) with "buyer" leads to Table 1.

Table 1: Cluster Evaluation with Buyers

Model	M_{sep}		M_p		M_g		M_{g*p}	
Buyer	no	yes	no	yes	no	yes	no	yes
Cluster 1	1331	244	1305	396	836	1112	1109	511
Cluster 2	143	1634	884	569	2130	214	111	539
Cluster 3	1550	145	194	1168	1076	949	1789	477
Cluster 4	1573	135	1286	417	981	401	498	478
Cluster 5	895	1086	1011	959	937	364	1291	159
Cluster 6	903	361	1715	96	435	565	1597	1441

Obviously, for M_{sep} Cluster 2 is mainly decomposed by buyers; Cluster 1, 3, and 4 mainly by non-buyers. Cluster 5 is also interesting since we have a considerably large amount of buyers in there as well. Basically, at this point the interpretation in an "ordinary" cluster analysis such as *k-means* would stop. With the proposed parametric survival approach we can examine (apart from probabilistic cluster assignment) interesting clusters in terms of their navigation behavior; i.e., the relative risk to leave a certain page depending on the dwell time. We can produce the hazard plots from two different perspectives: "group perspective", where for each cluster the survival/hazard functions over the single pages are plotted and "page perspective" where for each page these functions are plotted over the single clusters. We focus our analyses on the most appealing clusters 2, 3, and 5.

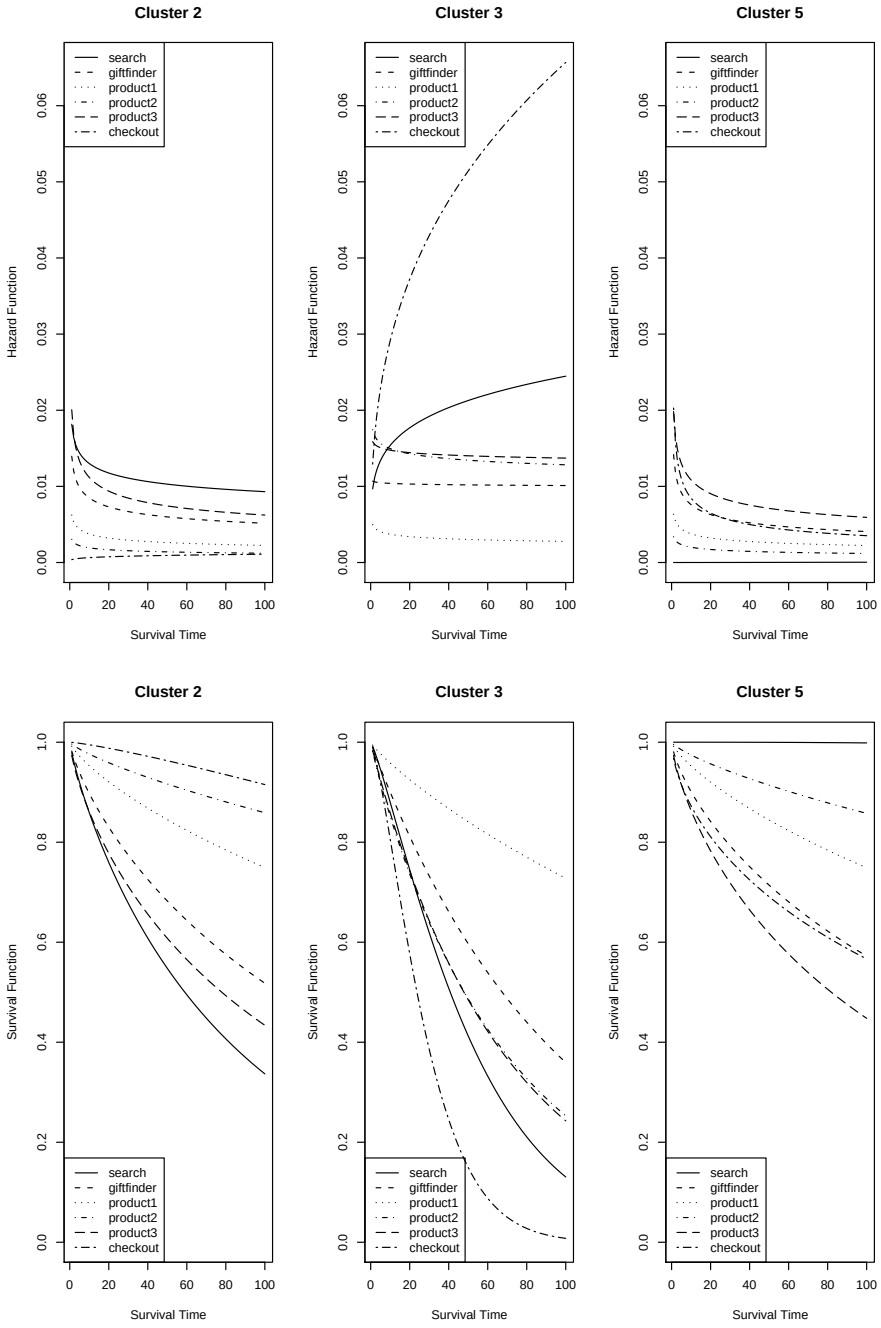


Figure 2: Hazard and Survival Function - Group Perspective

The survival plots in Figure 2 show the survival function by means of the dwell times. Hence, our cluster approach allows a detailed examination of the cluster behavior beyond common cluster centroids. In this section we will explain differences between various restrictive models. These become obvious in the hazard plots and therefore we will limit our further elaborations and interpretations on this plot type.

At first glance clusters 2 and 5 show a similar transition behavior. A remarkable difference can be found by examining the hazard for the "search" page: For Cluster 5 the hazard for this page-type is constantly close to 0. By inspecting the search-dwell times for this particular page we find lots of non-visits and a few rather large dwell times (> 600). In Cluster 2 the relative risk to leave this page is considerably large (throughoutly). However, for these two clusters all hazard functions (except Product 2 in Cluster 2) are decreasing. For Cluster 3 we have obvious differences in the hazards. They are increasing for the search page and the checkout area. These people (mostly non-buyers as seen before), if at all, visit the checkout page only for a short time. Let us represent these three crucial pages from the "page perspective" for across six clusters (see Figure 3).

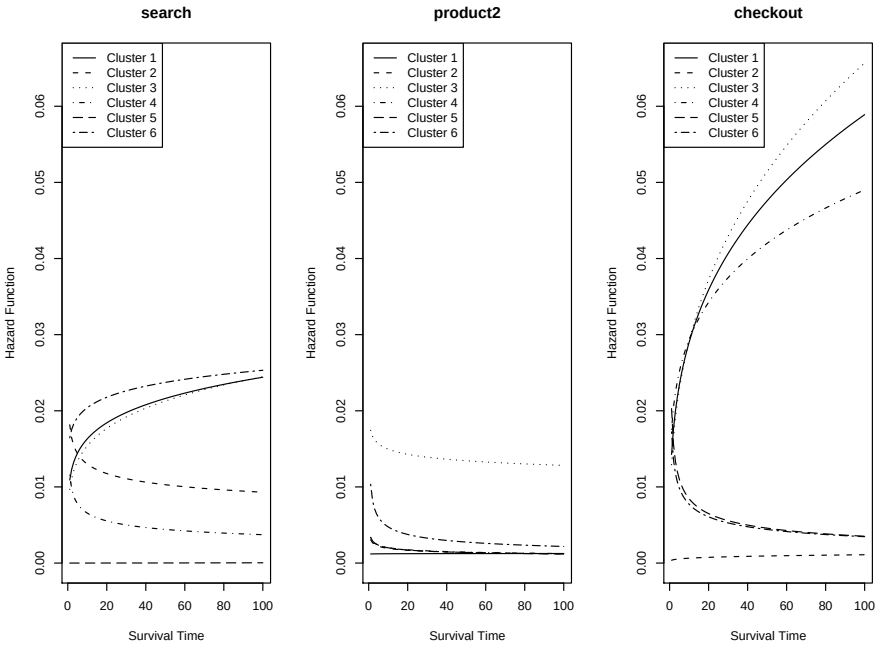


Figure 3: Hazard Function - Page Perspective

For the checkout area we have similar hazard patterns for clusters 1, 3, and 4. These are clusters with mostly non-buyers assigned. Note that for webshop providers dwell time models are helpful in detecting "anomalies" in the checkout area: Let us consider that a certain "non-buyer" cluster has large checkout dwell times and similar hazard patterns as "buyer" clusters on other pages. Obviously, members of the "non-buyer" cluster do not finish the checkout. This can be an advice that some parts of the checkout are not very user-friendly.

However, at this point the hazard behavior for the more restrictive WPHM is examined. All further plots are limited to the "group perspective". The hazard functions for M_p, which reflect proportionality over pages, are represented in Figure 4. Note that the cluster solution differs from M_{sep} as given in Table 1. The plot is limited to the most interesting clusters 3 (mainly buyers), 5 (buyer and non-buyer), and 6 (mainly non-buyers).

Each hazard function is montonically decreasing and furthermore, within each cluster the hazards are proportional to each other. Thus, from a practical point of view M_g, i.e., non-proportional

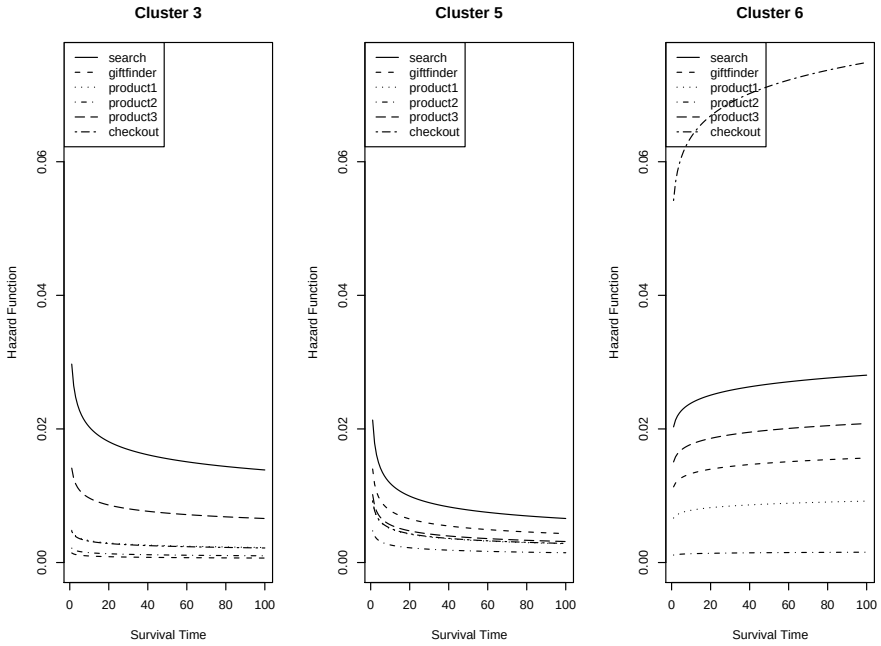


Figure 4: Hazard Function Page Proportionality

hazards for the pages within groups but proportional hazards for the groups within pages, is useful. The table for external buyer evaluation (Table 1) does not show such a clear separation between buyers and non-buyers over the clusters. The corresponding hazard functions for the clusters 1, 2, and 3 are given in Figure 5. By inspecting the dwell times for a certain page over the clusters it is obvious that they are proportional to each other.

For the model M_{g+p} the hazards are proportional in both directions which is typically too restrictive in web mining applications and thus it is not plotted here. This leads to the fact that the ranking of the hazard rates is the same over groups/pages. This can be relaxed by the interaction model M_{g*p} which is plotted in Figure 6 (Cluster 2, 4, 5).

Similar to M_g we do not have any clear buyer/non-buyer separation (see Table 1). Over the groups the Product 3 page has throughoutly the largest hazards. The hazard ranking is different: In Cluster 2 Product 3 is followed by search and gift finder; in Cluster 4, for instance, Product 1 is between Product 3 and gift finder. However, the proportionality restrictions hold in both directions. The interaction in M_{g*p} leads to different hazard orderings compared to M_{g+p} and is therefore less restrictive. In Cluster 5 all page hazards are very close to each other and hence these users show similar behavior on the site-areas.

5. Discussion

In this paper we presented a survival mixture approach for dwell time based session clustering. Based on the concept that dwell times correspond to survival times in medical statistics we adopted mixture Weibull modeling for the webshop data and enhanced the methodology by providing mixtures of proportional hazard models with incomplete data. The whole modeling approach can be regarded as a framework to cluster sessions with respect to different proportionality restrictions. Various types of the EM-algorithm can be used to estimate the parameters.

Practical problems can occur during the EM-iteration: First of all, it is possible that cluster consists of less than 2 observations. In this case the survival distribution cannot be estimated in

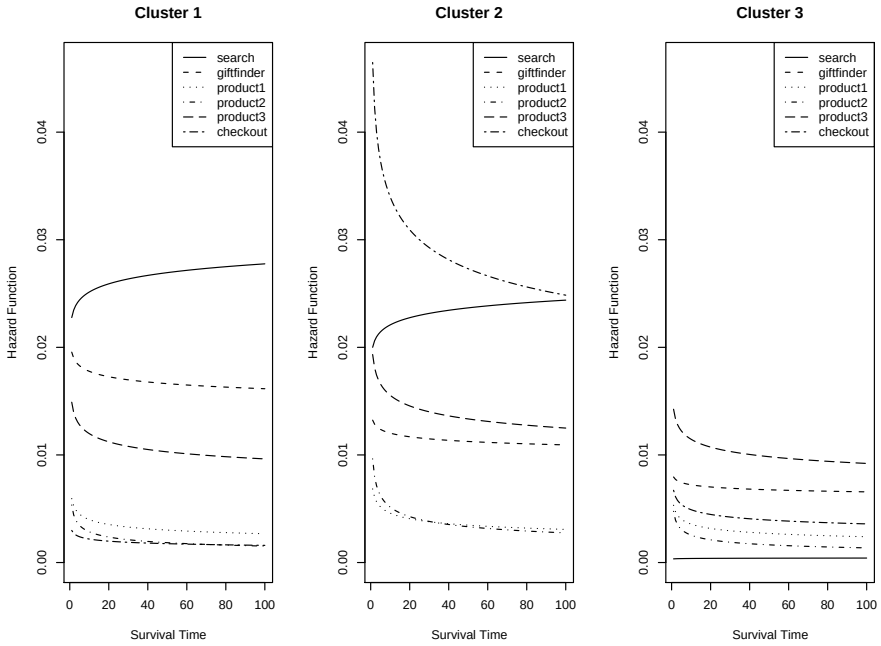


Figure 5: Hazard Function Group Proportionality

the E-step and the package `mixPHM` stops at this point. The user can re-run the analysis with the same K but with a different starting solution (e.g., from a precursory k-means clustering), decrease the number of components, or remove outliers, if any.

Furthermore, it can occur that within a cluster k a certain page p is not visited at all. This has to be taken into account when estimating $\widehat{\mathbf{\Lambda}}$ and $\widehat{\mathbf{\Gamma}}$. The corresponding elements $\hat{\lambda}_{k,p}$ and $\hat{\gamma}_{k,p}$ are missing and thus ignored in the likelihood computation.

In general, survival analysis provides censored data, i.e., individuals where the end-point of interest has not been observed ("right-censored"). The concept of censoring data can be applied straightforwardly to our dwell time case: For instance, if a user keeps a page open during the night. The corresponding dwell time is huge and not informative for cluster analysis. So far we excluded such dwell times but as a topic of future implementations the likelihood equations accounting for censored data will be established.

For exponential or Rayleigh distributions the hazard rates are constant. The Weibull assumption leads to hazard rates which are monotonic increasing or decreasing. For our example this assumption is reasonable and from an interpretational point of view it leads to valuable results for the provider. Due to the parsimonity of these models the results are easily understandable and they can be communicated straightforwardly. However, one may argue that these models are too restrictive. If the Weibull assumption does not hold, semi-parametric approaches such as the Cox-regression can be taken into account (Kim, Yun, and Dohi 2003). Alternatively, higher parametrized survival distributions such as the 3-parametric Hjorth distribution (Hjorth 1980) or beta-log-normal families (Walker and Stephens 1999) can be considered. In the latter case the hazard rates can increase and decrease over time. If the linearity in the PHM is doubted, also non-linear hazard models come into question as, e.g., given in Batchelor, Turner, and Firth (2007). However, with all these possible extensions the convergence of the EM algorithm has to be ensured theoretically.

Finally, the independence assumption can be doubted. As mentioned earlier, in order to apply our mixture approach the pages have to be categorized into proper site-areas. However, if such

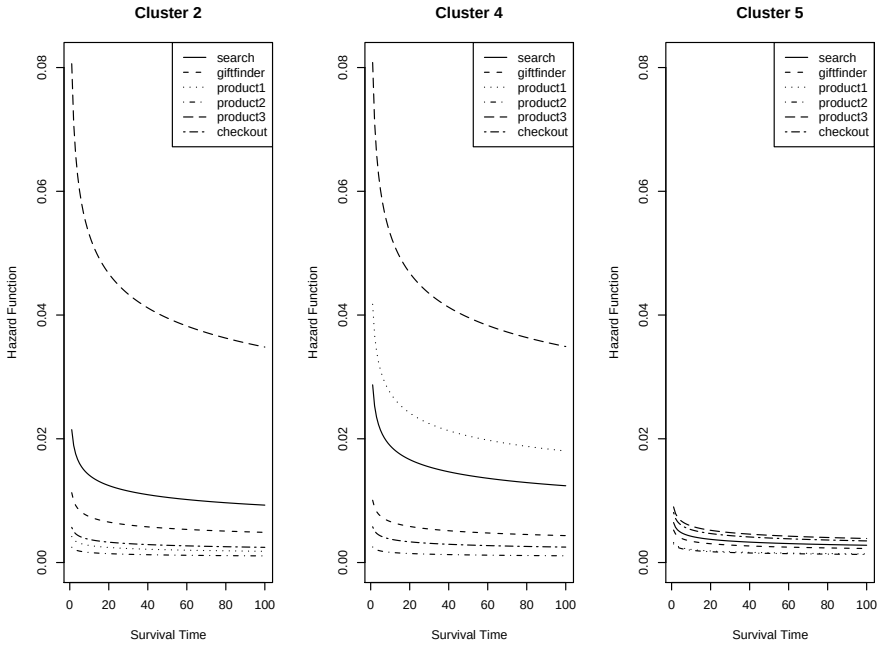


Figure 6: Hazard Function Group/Page Proportionality with Interactions

independent site segments cannot be achieved, the covariance structure between the pages has to be taken into account. This would correspond to a multivariate mixture Weibull distribution as for instance given in Patra and Dey (1999) and the mixtures of PHM as elaborated in Guo and Rodriguez (1992). From a practical point of view this is the most appealing extension of our approach and will be implemented in a subsequent version of the `mixPHM` package.

Further, the package allows for an explorative model selection (different number of clusters, different hazard restrictions) using the BIC criterion. Usually, the BICs decrease continuously with an increasing number of components. Thus, the aim is to find a reasonable cut point for k with the common trade-off between statistical goodness-of-fit and substantial interpretability. A BIC scree plot as provided in the package can be helpful.

For testing specific models several strategies can be considered. Graphically, since no censored observations are taken into account, the survival function from the underlying model can be plotted against the empirical survival function. If censored data are included, life-table or Kaplan-Meier estimators can be used. Testing for model fit can be accomplished by means of ordinary LR-tests if the models are nested, e.g., by testing M_p against M_{sep} for a fixed K. This example implies testing the proportional hazard assumption (Collett 2003). In the case of non-nested models (e.g., M_p with 3 components against M_{g+p} with 4 components) specific non-nested testing strategies have to be carried out.

Acknowledgments

The authors thank Viktor Krammer, Roland Kurzawa, and Johannes Sperlhofer at ec3 for ETL programming, data warehousing, data preparation, reporting, and descriptive statistical analyses on the underlying project.

References

Batchelor A, Turner HL, Firth D (2007). "Nonlinear discrete-time hazard models for the rate of first marriage." In J del Castillo, A Espinal, P Puig (eds.), "Proceedings of the 22nd International Workshop on Statistical Modelling," pp. 99–102. IDESCAT, Barcelona.

Cadez I, Heckerman D, Meek C, Smyth P, White S (2001). "Visualization of navigation patterns on a web site using model based clustering." *Data Mining and Knowledge Discovery*, **7**, 399–424.

Celeaux G, Govaert G (1992). "A classification EM algorithm for clustering and two stochastic versions." *Computational Statistics and Data Analysis*, **14**, 315–332.

Collett D (2003). *Modelling Survival Data in Medical Research.* Chapman & Hall/CRC, Boca Raton, FL, 2nd edition.

Cox DR (1972). "Regression models and life-tables (with discussion)." *Journal of the Royal Statistical Society, Series B*, **74**, 187–220.

Cox DR, Oakes D (1984). *Analysis of Survival Data.* CRC Press, Boca Raton, FL.

Dempster AP, Laird NM, Rubin DB (1977). "Maximum likelihood from incomplete data via the EM-algorithm." *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Guo G, Rodriguez G (1992). "Estimating a Multivariate Proportional Hazards Model for Clustered Data Using the EM Algorithm, with an Application to Child Survival in Guatemala." *Journal of the American Statistical Association*, **87**, 969–976.

Heckman J, Singer B (1984). "A method for minimizing the impact of distributional assumptions in econometric models for duration data." *Econometrica*, **58**, 453–473.

Hjorth U (1980). "A reliability distribution with increasing, decreasing, constant and bathtub-shaped failure rates." *Technometrics*, **22**, 99–109.

Honoré BE (1990). "Simple estimation of a duration model with unobserved heterogeneity." *Econometrica*, **58**, 453–473.

Ishwaran H (1996). "Identifiability and rates of estimation for scale parameters in location mixture models." *The Annals of Statistics*, **24**, 1560–1571.

Jewell NP (1982). "Mixtures of exponential distributions." *The Annals of Statistics*, **10**, 479–484.

Kalbfleisch JD, Prentice RL (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York, NY.

Kim JW, Yun WY, Dohi T (2003). "Estimating the mixture of proportional hazards model with incomplete failure data." *Journal of Quality in Maintenance Engineering*, **9**, 265–278.

Mair P, Hudec M (2007). *mixPHM: R package for mixtures of proportional hazard models.* URL: `http://cran.R-project.org`.

Mair P, Hudec M (2008). "Analysis of dwell times in web usage mining." In "Proceedings of the 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (in press)," Springer, New York.

McLachlan GJ, Krishnan T (1997). *The EM Algorithm and Extensions.* Wiley, New York, NY.

McLachlan GJ, Peel D (2000). *Finite Mixture Models.* Wiley, New York, NY.

Montgomery AL, Li S, Srinivasan K, Liechty JC (2004). "Modeling online browsing and path analysis using clickstream data." *Marketing Science*, **23**, 579–595.

Park Y, Fader PS (2004). "Modeling browsing behavior at multiple websites." *Marketing Science*, **23**, 280–303.

Patra K, Dey DK (1999). "A multivariate mixture of Weibull distributions in reliability modeling." *Statistics & Probability Letters*, **45**, 225–235.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL: `http://www.R-project.org`. ISBN 3-900051-07-0.

Ramon J, Albert G, Baxter LA (1995). "Applications of the EM algorithm to the analysis of life-length data." *Applied Statistics*, **44**, 323–341.

Smyth P (1999). "Probabilistic model-based clustering of multivariate and sequential data." In D Heckerman, J Whittaker (eds.), "Proceedings of Seventh International Workshop on Artificial Intelligence and Statistics," pp. 299–304. Morgan Kaufman, San Francisco, CA.

Walker SG, Stephens DA (1999). "A multivariate family of distributions on $(0, \infty)^p$." *Biometrika*, **86**, 703–709.

Ypma A, Heskes T (2002). "Automatic categorization of web pages and user clustering with mixtures of hidden Markov models." In OR Zaiane, J Srivastava, M Spiliopoulou, B Masand (eds.), "Proceedings of the 4th International Workshop on Mining Web Data for Discovering Usage Patterns and User Profiles," pp. 35–49. Springer, New York, NY.