**ORIGINAL PAPER**

Check for
updates

# Clustering method for censored and collinear survival data

**Silvia Liverani[1,2]** · **Lucy Leigh[3,4]** · **Irene L. Hudson[5]** · **Julie E. Byles[4]**

## Abstract

In this paper we propose a Dirichlet process mixture model for censored survival data with covariates. This model is suitable in two scenarios. First, this method can be used to identify clusters determined by both the censored survival data and the predictors. Second, this method is suitable for highly correlated predictors, in cases when the usual survival models cannot be implemented because they would be unstable due to multicollinearity. The Dirichlet process mixture model links a response vector to covariate data through cluster membership and in this paper this model is extended for mixtures of Weibull distributions, which can be used to model survival times and also allow for censoring. We propose two variants of this model, one with a shape parameter common to all clusters (referred to as a global parameter) for the Weibull distributions and one with a cluster-specific shape parameter. The first satisfies the proportional hazard assumption, while the latter is very flexible, as it has the advantage of allowing estimation of the survival curve whether or not the proportional hazards assumption is satisfied. We present a simulation study and, to demonstrate the applicability of the method in practice, a real application to sleep surveys in older women from The Australian Longitudinal Study on Women's Health. The method developed in the paper is available in the R package PReMiuM.

**Keywords** Clustering · Bayesian non-parametric · Sleep data

✉ Silvia Liverani
  liveranis@gmail.com

1   School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, UK

2   The Alan Turing Institute, The British Library, London NW12DB, UK

3   Clinical Research Design, IT and Statistical Support Unit, Hunter Medical Research Institute, New Lambton Heights, NSW 2305, Australia

4   Research Centre for Generational Health and Ageing, University of Newcastle, University Drive, Callaghan, NSW 2308, Australia

5   Department of Mathematical Sciences, College of Science, Engineering and Health, Royal Melbourne Institute of Technology (RMIT), City Campus, 124 La Trobe Street,, Melbourne VIC 3000, Australia

🛆 Springer

# 1 Introduction

We propose a Dirichlet process mixture model for censored survival data with covariates. This model is most useful in two situations.

First of all, this method can be used to identify clusters determined by censored survival data and explanatory variables. The idea of linking a response vector to covariate data through cluster membership was proposed initially by several authors including Dunson et al. (2008), Bigelow and Dunson (2009), Molitor et al. (2010), Papathomas et al. (2011), and Molitor et al. (2011). We will focus on the latter of these articles, which refers to this idea as *profile regression*, where a Dirichlet mixture model is used for inference on the clusters. This model was implemented in an R package by Liverani et al. (2015) and it has been employed in a variety of fields (Molitor et al. 2014), including genetics (Papathomas et al. 2012), environmental epidemiology (Papathomas et al. 2011; Pirani et al. 2015; Coker et al. 2016; Liverani et al. 2016) and occupational epidemiology (Hastie et al. 2013; Mattei et al. 2016). In this paper we extend this model to survival outcomes with censoring.

Second, the proposed method is suitable when the explanatory variables are multicollinear. Multicollinearity, or collinearity, is the existence of near-linear relationships among the explanatory variables. The high correlation between explanatory variables can create inaccurate, or unstable, estimates of the regression coefficients, inflate the standard errors, deflate the partial t-tests, give false, nonsignificant, p-values, and degrade the predictability of the model. Hence, one of the first steps in a regression analysis is to determine if multicollinearity is present. Our proposed method is stable when highly correlated predictors are included in the model, making it a powerful tool to explore survival datasets with highly correlated predictors.

The model that we propose is essentially a mixture of Weibull distributions and distributions suitable for the covariates, non-parametrically linking the response and the predictors through cluster membership. Modelling independently the response and the covariates is the idea underpinning profile regression as an exploratory method in the presence of collinearity in the covariates. This modelling choice allows the exploration of the complex relationship between the response and the covariates. Although the response and the covariates are modelled independently, this clustering method can uncover linear and non-linear relationships between covariates and response.

This model includes some cluster specific parameters, which characterise the clusters, and some global parameters, which are shared by all clusters. The Weibull distributions, with cluster-specific scale parameters, can be used to model survival times and also allow for censoring. We propose two models with Weibull distributions for the response, one with a global shape parameter for the Weibull distributions and one with a cluster-specific shape parameter. The first model satisfies the proportional hazard assumption, which allows for comparisons between clusters. On the other hand, the latter model has the advantage of allowing the estimation of the survival curve without having to satisfy the proportional hazards assumption. Therefore, it is a very flexible model. Suitable distributions for the covariates are the Normal distribution in the case of continuous explanatory variables and the multinomial distribution in the case of categorical explanatory variables.

Kottas ([2006](#)) did important early work on the Weibull Dirichlet process mixture model for unknown survival distributions, although their model was limited to estimating the survival distribution component only and did not extend to regression. In contrast, our proposed method links the survival outcome to a multivariate profile, and estimates hazard ratios (in the case where a proportional hazards assumption is satisfied) and median survival times. Moreover, the Weibull Dirichlet process mixture by Kottas ([2006](#)) involves mixing on both the shape and scale parameters of the Weibull kernel, while we also propose and discuss the reduced model satisfying the desirable assumption of proportional hazards. Another contribution of our paper is the computation of the posterior predictive distribution of survival time to provide interpretable results. Finally, our methods are readily available in the R package PReMiuM using advanced state-of-the-art MCMC algorithms.

In this paper, we apply our model to the analysis of sleep data based on a unique cohort of very old women from The Australian Longitudinal Study on Women's Health (ALSWH). We are interested in learning about the relationship between sleep difficulty and survival in an Australian cohort of old women (Leigh et al. [2016b](#)). Due to the fact that difficulty in sleeping may be related to additional factors which also affect survival (for example, Body Mass Index (BMI), comorbidity, sleep medication use, physical functioning and vitality, mental health), it is also of interest to model the joint effects of sleep difficulty and these additional covariates on survival, via profile regression. Previous analyses (Leigh et al. [2015](#), [2016a](#)) utilised latent class analysis (LCA) to identify longitudinal patterns (profiles) of sleep difficulty, and then utilised these classes to predict survival, adjusted for various other factors, as well as the interaction between the sleep classes and disease count. In the present paper, profiles are based on the additional covariates as well as sleep difficulty, and thus may better capture the complex interactions between all covariates of interest. Moreover, only a single model fit is required rather than a procedure in steps, which might be unable to model appropriately certain features of the data.

In Sect. [2](#) we introduce the formulation of the Dirichlet process mixture model and profile regression. In Sects. [3](#) and [4](#) we propose the two new models for censored survival data with global and cluster-specific shape parameters. In Sect. [5](#) we provide a method for the computation of hazard ratios, expected survival time and predictions. In Sect. [6](#) we report the results on simulated data and in Sect. [7](#) the results on the ALSWH dataset. Some concluding remarks are given in Sect. [8](#).

## 2 Profile regression

Profile regression is a Dirichlet process mixture model where the response variable and the covariates are modelled jointly (Molitor et al. [2010](#); Liverani et al. [2015](#)).

The Dirichlet process (DP) is a stochastic process used in Bayesian nonparametric models, particularly in Dirichlet process mixture models. It is a distribution over distributions, so each draw from a Dirichlet process is itself a distribution. For a random distribution $G$ to be distributed according to a DP, its marginal distributions have to be Dirichlet distributed, which is the reason for the name Dirichlet of this process. Specifically, let $H$ be a distribution over $\Theta$ and $\alpha$ be a positive real number. Then

for any finite measurable partition $A_1, \ldots, A_r$ of $\Theta$, the vector $(G(A_1), \ldots, G(A_r))$ is random since $G$ is random. We say $G$ is Dirichlet process distributed with base distribution $H$ and concentration parameter $\alpha$, written $G \sim DP(\alpha, H)$, if

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r))$$

for every finite measurable partition $A_1, \ldots, A_r$ of $\Theta$. The draws from a DP satisfy a discreteness property which also implies a clustering property. The discreteness and clustering properties of the DP play crucial roles in the use of DPs for clustering via DP mixture models, as described in Teh (2011). The nonparametric nature of the Dirichlet process translates to mixture models with a countably infinite number of components. We model a set of observations $\{y_1, \ldots, y_n\}$ using a set of latent parameters $\{\theta_1, \ldots, \theta_n\}$. Each $\theta_i$ is drawn independently and identically from $G$, while each $y_i$ has distribution $F(\theta_i)$. Because $G$ is discrete, multiple $\theta_i$'s can take on the same value simultaneously, and the above model can be seen as a mixture model, where $y_i$'s with the same value of $\theta_i$ belong to the same cluster.

Profile regression is a generalisation of the DP mixture model, where the induced mixture model is a mixture of two distributions, one for the response vector $y$ and one for the covariate data $\mathbf{x}$. In particular, we define response data $y_i$ and covariate data $\mathbf{x}_i$ for each individual $i$ with $i = 1, \ldots, n$. There is also the possibility to include additional data, $\mathbf{w}_i$ for each individual, which we will refer to as fixed effects. The fixed effects are constrained to only have a global (i.e., non-cluster specific) effect on the response $y_i$ and the functional relationship between the response and the fixed effects is discussed below for specific response models. The mixture model is then given by

$$f(\mathbf{x}_i, y_i | \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\beta}, z, \boldsymbol{w}_i) = \sum_{c=1}^{\infty} \psi_c f_x(\mathbf{x}_i | z_i = c, \boldsymbol{\phi}_c) f_y(y_i | z_i = c, \theta_c, \boldsymbol{\lambda}, \mathbf{w}_i) \quad (1)$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})$ is the $P$-dimensional covariate profile and $\boldsymbol{z} = (z_1, \ldots, z_n)$ with $z_i = c$ is the allocation variable indicating the cluster to which individual $i$ belongs. The parameter vectors $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are the cluster specific parameters and are defined in more detail below. The parameter vector $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots)$ are the cluster weights and $\boldsymbol{\lambda}$ are the global parameters linking the fixed effects to the response variable. An active cluster is a cluster which contains at least one observation. There are an infinite number of clusters in this model, though a finite data set only exhibits a finite number of active clusters, which are inferred from the data.

The likelihoods $f_y$ and $f_x$ depend upon the choice of response and covariate model, respectively. The covariate model is different depending on the data. For continuous data, we assume a mixture of Gaussian distributions. Under this setting for each cluster $c$, the cluster specific parameters are given by $\boldsymbol{\phi}_c = (\mu_c, \Sigma_c)$, where $\mu_c$ is a mean vector and $\Sigma_c$ is a covariance matrix. Under this setting, it follows that

$$f_x(\mathbf{x}_i | z_i = c, \boldsymbol{\phi}_c) = (2\pi)^{-\frac{J}{2}} |\Sigma_c|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \mu_c)^{\top} \Sigma_c^{-1} (\mathbf{x}_i - \mu_c) \right\}. \quad (2)$$

For discrete variables, where for each individual $i$, $\mathbf{x}_i$ is a vector of $J$ locally independent discrete categorical random variables, where the number of categories for covariate $j$ is $K_j$, for $j = 1, 2, \ldots, J$. Then we can write $\boldsymbol{\phi}_c = \Phi_c = (\Phi_{c,1}, \Phi_{c,2} \ldots, \Phi_{c,J})$ with $\Phi_{c,j} = (\phi_{c,j,1}, \phi_{c,j,2}, \ldots, \phi_{c,j,K_j})$ and

$$f_x(\mathbf{x}_i | z_i = c, \Phi_c) = \prod_{j=1}^{J} \phi_{c,j,x_{i,j}}. \tag{3}$$

Similarly, for the response model we implement models which are suitable for the data under study. One simple case is a continuous response variable, in which case the likelihood for the response model is given by

$$f_y(y_i | z_i = c, \theta_c, \boldsymbol{\lambda}, \mathbf{w}_i) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left\{ -\frac{1}{2\sigma_y^2}(Y_i - \mu_i)^2 \right\}, \tag{4}$$

where $\mu_i = \theta_c + \boldsymbol{\beta}^\top \mathbf{w}_i$ and $\boldsymbol{\lambda} = (\boldsymbol{\beta}, \sigma_y^2)$. For both $f_y$ and $f_x$ we can also make other modelling choices, like a binary response model, a categorical response model, Poisson or Binomial mixtures for count data. Liverani et al. (2016) also propose an extension of this response model to account for spatial correlation. In this paper we propose a new response model, for censored survival data.

Profile regression as described above is implemented in the R package PReMiuM (Liverani et al. 2015), along with a range of prior distributions. Inference is made in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods. Hastie et al. (2015) provide details on assessing lack of convergence for these models. Additional features are available in the R package, such as two methods for variable selection, which allow us to determine which covariates actively drive the mixture components, and which share characteristics common to all components. One of these variable selection methods is based on the work by Chung and Dunson (2009), a cluster-specific selection approach which is also applied on the ALSWH sleep data in Sect. 7. Each cluster $c$ has an associated vector $\xi_c = (\xi_{c,1}, \xi_{c,2}, \ldots, \xi_{c,J})$, where $\xi_{c,j}$ is a binary random variable that determines whether covariate $j$ is important to cluster $c$. For discrete covariates, we can then define the new composite parameters,

$$\phi^*_{c,j,k} := \xi_{c,j}\phi_{c,j,k} + (1 - \xi_{c,j})\phi_{0,j,k} = \left(\phi_{c,j,k}\right)^{\xi_{c,j}} \left(\phi_{0,j,k}\right)^{(1-\xi_{c,j})} \tag{5}$$

which replace the cluster specific parameters for discrete covariates defined above. We assume that, given $\rho_j$, the $\xi_{c,j}, c = 1, \ldots, C$, are independent Bernoulli variables with $\xi_{c,j} \sim \text{Bernoulli}(\rho_j)$. To induce variable selection, we consider a sparsity inducing prior for $\rho_j$ with an atom at zero, so that

$$\rho_j \sim 1_{\{w_j=0\}}\delta_0(\rho_j) + 1_{\{w_j=1\}}\text{Beta}(\alpha_\rho, \beta_\rho), \tag{6}$$

where $w_j \sim$ Bernoulli($p_w$). By examining the posterior distribution of $\rho_j$ we can ascertain the extent of the contribution of variable $j$ to the clustering: if it has mostly mass around zero, it is unlikely to be contributing significantly to the clustering.

The MCMC produces a rich posterior output, with a partition of the observations provided at each iteration. It is therefore necessary to infer a representative partition, as an effective way to convey the output of the clustering algorithm. It is also of interest to assess the uncertainty associated with subgroups of this best partition. Moreover, due to the problem of 'label switching', i.e the labels associated with each cluster change during the MCMC iterations, we can not simply assign each observation to the cluster that maximizes the average posterior probability. One solution which has proved useful is to summarise the MCMC output in a dissimilarity matrix, where at each iteration of the sample, we record pairwise cluster membership and construct a score matrix. Averaging these matrices over the whole MCMC run leads to a similarity matrix S, which can then be used to identify an optimal partition. Post-processing methods are also available in the R package PReMiuM and discussed in detail by Liverani et al. (2015). Molitor et al. (2010) include further discussion on the motivation and justification of profile regression models.

## 3 Survival response Weibull with global shape parameter

We extend the profile regression model described in Sect. 2 for survival data with censoring, using a mixture of Weibull distributions. In this section, we develop the model with a global shape parameter for the Weibull distribution.

For survival data, with a survival or censoring time and a censoring indicator, we have

$$f_y(y_i|z_i = c, \theta_c, \boldsymbol{\lambda}, \mathbf{w}_i) = h(y_i|z_i = c, \theta_c, \nu, \boldsymbol{\beta}, \mathbf{w}_i)^{d_i} S(y_i|z_i = c, \theta_c, \nu, \boldsymbol{\beta}, W_i) \quad (7)$$

where $h$ is the hazard function, $S$ is the survival function, $\boldsymbol{\lambda} = (\nu, \boldsymbol{\beta})$ are the global parameters and $y$ is the lifetime of an individual. The censoring indicator $d_i$ is defined as follows.

$$d_i = \begin{cases} 0 & \text{if the individual is censored or} \\ 1 & \text{if the individual experiences the event of interest} \end{cases} \quad (8)$$

with $d = \sum_{i=1}^n d_i$. Survival time has a Weibull distribution if its survival distribution is given by

$$S(y_i|z_i = c, \theta_c, \nu, \boldsymbol{\beta}, \mathbf{w}_i) = f(y > y_i|\theta_{z_i}, \nu, \boldsymbol{\beta}, \mathbf{w}_i) = \exp\left(-\gamma_{z_i} y_i^\nu\right) \quad (9)$$

and its hazard function is given by

$$h(y_i|z_i = c, \theta_c, \nu, \boldsymbol{\beta}, \mathbf{w}_i) = \nu \gamma_{z_i} y_i^{\nu-1} \quad (10)$$

with link function $\gamma_{z_i} = \exp\left(\theta_{z_i} + \boldsymbol{\beta}^T \mathbf{w}_i\right)$. For this model the baseline risk is constant. When $\nu > 1$ the hazard rate increases as time increases, it is constant for $\nu = 1$ and the hazard rate decreases for $\nu < 1$.

The likelihood is given by

$$f_y(y|.) = \prod_{i=1}^{n} h(y_i|z_i = c, \theta_c, \nu, \boldsymbol{\beta}, \mathbf{w}_i)^{d_i} S(y_i|z_i = c, \theta_c, \nu, \boldsymbol{\beta}, W_i) \quad (11)$$

$$= \nu^d \left(\prod_{i=1}^{n} \gamma_{z_i}^{d_i}\right) \exp\left(-\sum_{i=1}^{n} \gamma_{z_i} y_i^{\nu}\right) \prod_{i=1}^{n} \left(y_i^{\nu-1}\right)^{d_i}. \quad (12)$$

Therefore, the conditional distribution of $\nu$ is given by

$$f(\nu|.) \propto \prod_{i=1}^{n} h(y_i|z_i = c, \theta_c, \nu, \boldsymbol{\beta}, \mathbf{w}_i)^{d_i} S(y_i|z_i = c, \theta_c, \nu, \boldsymbol{\beta}, W_i)\pi_\nu(\nu) \quad (13)$$

$$\propto \nu^d \exp\left(-\sum_{i=1}^{n} \gamma_{z_i} y_i^{\nu}\right) \prod_{i=1}^{n} \left(y_i^{\nu-1}\right)^{d_i} \pi_\nu(\nu) \quad (14)$$

where $\pi_\nu(\nu)$ is the log-concave prior distribution of $\nu$. It can be shown that

$$\frac{\partial^2 \log f(\nu|.)}{\partial \nu^2} < 0$$

which is satisfied if and only if $f(\nu|.)$ is log-concave (Borzadaran and Borzadaran 2011). Given the log concavity of $f(\nu|.)$, we can use an adaptive rejection sampling algorithm to sample from the posterior distribution of $\nu$ (Gilks and Wild 1992). We set the prior distribution for $\nu$, $\pi_\nu(\nu)$, to be a Gamma distribution with parameters $a_\nu$ and $b_\nu$, so we require that $a_\nu \geq 1$ to ensure the log-concavity of $\pi_\nu(\nu)$.

To implement the adaptive rejection sampler (Gilks and Wild 1992), we require the logarithm of a function proportional to the distribution of interest and its derivative. This function and its derivative are given by the following,

$$\log f(\nu|.) \propto d \log \nu - \sum_{i=1}^{n} \gamma_{z_i} y_i^{\nu} + \nu \sum_{i=1}^{n} d_i \log y_i + (a_\nu - 1) \log \nu - b_\nu \nu \quad (15)$$

and

$$\frac{\partial \log f(\nu|.)}{\partial \nu} \propto \frac{d}{\nu} - \sum_{i=1}^{n} \gamma_{z_i} y_i^{\nu} \log y_i + \sum_{i=1}^{n} d_i \log y_i + \frac{a_\nu - 1}{\nu} - b_\nu. \quad (16)$$

The model developed in this section has a global shape parameter for each Weibull distribution in the mixture. The advantage of this model is that the assumption of proportional hazards holds and we can compute hazard ratios between different clusters.

However, for the cases where the assumption of proportional hazards is untenable, we develop a more flexible model, in the Sect. 4, with cluster-specific shape parameters.

## 4 Survival response Weibull with cluster-specific shape parameter

Here we propose a mixture of Weibull distributions with cluster-specific scale and shape parameters. This model is more flexible than the model proposed in Sect. 3 because it does not require the assumption of proportional hazards to hold. In this model the shape parameters of the Weibull distributions are now a vector $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots)$ of cluster-specific shape parameters. Therefore, the components of the mixture of Weibull distributions take the following form,

$$f_y(y_i|z_i = c, \theta_c, \nu_c, \boldsymbol{\lambda}, \mathbf{w}_i) = h(y_i|z_i = c, \theta_c, \nu_c, \boldsymbol{\beta}, \mathbf{w}_i)^{d_i} S(y_i|z_i = c, \theta_c, \nu_c, \boldsymbol{\beta}, W_i) \tag{17}$$

where $h$ is the hazard function, $S$ is the survival function, $\boldsymbol{\lambda} = \boldsymbol{\beta}$ are the global parameters, $\theta_c$, $\nu_c$ and $\phi_c$ are the cluster-specific parameters and $y_i$ is the lifetime of an individual. It follows that the survival time $Y$ has a Weibull distribution if its survival distribution is now given by

$$S(y_i|z_i = c, \theta_c, \nu_{z_i}, \boldsymbol{\beta}, \mathbf{w}_i) = f(y > y_i|\theta_{z_i}, \nu_{z_i}, \boldsymbol{\beta}, \mathbf{w}_i) = \exp\left(-\gamma_{z_i} y_i^{\nu_{z_i}}\right) \tag{18}$$

and its hazard function is as follows

$$h(y_i|z_i = c, \theta_c, \nu_{z_i}, \boldsymbol{\beta}, \mathbf{w}_i) = \nu_{z_i} \gamma_{z_i} y_i^{\nu_{z_i}-1} \tag{19}$$

with link function $\gamma_{z_i} = \exp\left(\theta_{z_i} + \boldsymbol{\beta}^T \mathbf{w}_i\right)$. For this model the baseline risk is constant. The loglikelihood is given by

$$\log f_y(y|.) = \sum_{i=1}^{n} \log\left(\left(\nu_{z_i} \gamma_{z_i} y_i^{\nu_{z_i}-1}\right)^{d_i} \exp\left(-\gamma_{z_i} y_i^{\nu}\right)\right) \tag{20}$$

$$= \sum_{i=1}^{n} \left(d_i \left(\log \nu_{z_i} + \log \gamma_{z_i} + (\nu_{z_i} - 1) \log y_i\right) - \gamma_{z_i} y_i^{\nu}\right). \tag{21}$$

The conditional distribution of $\nu_c$ depends only on the data in cluster $c$. We define the indicator $d_{lc}$ for the censored data in cluster $c$.

$$d_{lc} = \begin{cases} 0 & \text{if the individual } l \text{ in cluster } c \text{ is censored} \\ 1 & \text{if the individual } l \text{ in cluster } c \text{ experiences the event.} \end{cases} \tag{22}$$

with $d_c = \sum_{l=1}^{n_c} d_{lc}$ and $\sum_{c=1}^{C} n_c = n$. It follows that the conditional distribution of $v_c$, for $z_i = c$, is given by

$$\log f(v_c|.) \propto \sum_{l=1}^{n_c} \left( d_{lc} \left( \log v_c + \log \gamma_{zl} + (v_c - 1) \log y_l \right) - \gamma_{zl} y_i^{v_c} \right) + \log \pi_{v_c}(v_c) \quad (23)$$

$$\propto \left( \log v_c + \log \gamma_{zl} \right) d_c + (v_c - 1) \sum_{l=1}^{n_c} d_{lc} \log y_l - \sum_{l=1}^{n_c} \gamma_{zl} y_i^{v_c} + \log \pi_{v_c}(v_c)$$

$$(24)$$

where $\pi_v(v_c)$ is the log-concave prior distribution of $v_c$. It can be easily shown that

$$\frac{\partial^2 \log f(v|.)}{\partial v^2} < 0$$

which, as before, is satisfied if and only if $f(v|.)$ is log-concave. Given the log concavity of $f(v|.)$, as for the case of the global shape parameter, it follows that we can use an adaptive rejection sampling algorithm for $v$. We set the prior distribution for each shape parameter $v_c$ to be a Gamma distribution with parameter $a_v$ and $b_v$, and require that $a_v \geq 1$ to ensure the log-concavity of $\pi_v(v_c)$. As before, to implement the adaptive rejection sampler, we require the logarithm of a function proportional to the distribution of interest and its derivative, which are given by

$$\log f(v_c|.) \propto d_c \log v_c + v_c \sum_{l=1}^{n_c} d_{lc} \log y_l - \sum_{l=1}^{n_c} \gamma_{zl} y_l^{v_c} + (a_v - 1) \log v_c - b_v v_c \quad (25)$$

and

$$\frac{\partial \log f(v_c|.)}{\partial v_c} \propto \frac{d_c}{v_c} - \sum_{l=1}^{n_c} d_{lc} \log y_l + \sum_{l=1}^{n_c} \gamma_{zl} y_l^{v_c} \log y_l + \frac{a_v - 1}{v_c} - b_v. \quad (26)$$

This mixture model with cluster-specific shape parameters can fit the data well in each cluster, but the assumption of proportional hazards does not hold. The additional challenge is how to compare observations in different clusters informatively. A proposal for this is given in the following section.

## 5 Computing and interpreting the hazard ratios and the expected survival time

The main inferential objective is to compare the clusters identified by profile regression. This is straightforward when the shape parameter is global, but cluster comparisons require careful consideration when the shape parameter is cluster specific. An alternative approach is to compare the clusters using the predicted survival time for individuals that belong to different clusters.

When there is a global shape parameter $\nu$ we can easily compute hazard ratios. The ratio of the hazard functions of two different clusters, with all fixed effects $\mathbf{w}_i$ constant, is given by

$$\frac{h(y_i|c_1)}{h(y_i|c_2)} = \frac{\nu\gamma_{c_1}y_i^{\nu-1}}{\nu\gamma_{c_2}y_i^{\nu-1}} = \frac{\nu\exp(\theta_{c_1} + \boldsymbol{\beta}^T\mathbf{w}_i)y_i^{\nu-1}}{\nu\exp(\theta_{c_2} + \boldsymbol{\beta}^T\mathbf{w}_i)y_i^{\nu-1}} = \frac{\exp(\theta_{c_1})}{\exp(\theta_{c_2})} = \exp(\theta_{c_1} - \theta_{c_2}).$$
(27)

Moreover, we can compute the hazard ratios for the fixed effects. The ratio of the hazard functions of two different values of $w_j$ in cluster $c_k$ is given by

$$\frac{h(y_i|c_k, w_j = x_1)}{h(y_i|c_k, w_j = x_2)} = \frac{\nu\exp(\theta_{c_k} + \beta_1 w_1 + \cdots + \beta_p w_p)y_i^{\nu-1}}{\nu\exp(\theta_{c_k} + \beta_1 w_1 + \cdots + \beta_p w_p)y_i^{\nu-1}} = \exp(\beta_j(x_1 - x_2)).$$
(28)

The fixed effects $\beta_j$ are global parameters, so they take the same value within each cluster. Therefore, we can write

$$\frac{h(y_i|c_k, w_j = x_1)}{h(y_i|c_k, w_j = x_2)} = \frac{h(y_i'|c_l, w_j = x_1)}{h(y_i'|c_l, w_j = x_2)} = \exp(\beta_j(x_1 - x_2)) \qquad (29)$$

for any $l$ and $k$. These ratios are constant proportions that depend only on the covariate $w_j$ and not on time. If $x_1 = x_2 + 1$ then the hazard ratio simplifies to $\exp(\beta_j)$.

In the case of cluster-specific shape parameter $\nu$, first we check whether we can assume proportional hazards. The assumption that the proportional hazards stay constant over time can be inspected by looking at a graph of the logarithm of the estimated cumulative hazard function. This plot is also known as a log-log survival plot. The proportional hazard assumption is evidenced by the difference between the logarithms of the hazards for any two clusters not changing over time, or equally by the difference between the logarithms of the cumulative hazard functions being constant. If proportional hazards are a sensible assumption, we can compute the hazards as above. If the hypothesis of proportional hazards is not tenable, we can interpret the results by computing the mean survival time. The mean survival time for cluster $c$ is given by

$$E(Y) = \gamma_c^{-1/\nu_c}\Gamma(1 + 1/\nu_c) \qquad (30)$$

where $\Gamma(.)$ represents the Gamma function.

## 5.1 Predictions

Posterior predictive distributions are computed for the survival time and the hazard ratios. At each sweep, the allocation of a predictive profile to a cluster $c$ is sampled from the mixture weights, according to the covariates $x_{pred}$ of the predictive profile. These draws give us a posterior predictive distribution for the $\hat{\theta}_s$, which is the predicted value of $\theta_{z_s}$ for the predictive profile $s$ at the $r$-th iteration of the MCMC.

**Table 1** Values of the parameter $\theta$ and $\phi$ for each cluster

| Cluster | $\theta_c$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---------|-----------|--------|--------|-------------|-------------|--------------------------|
| 1 | $-6.5$ | 0.8, 0.2 | 0.2, 0.8 | 0.1, 0.1, 0.8 | 0.1, 0.8, 0.1 | 0.25, 0.25, 0.25, 0.25 |
| 2 | $-4$ | 0.2, 0.8 | 0.2, 0.8 | 0.1, 0.1, 0.8 | 0.1, 0.8, 0.1 | 0.1, 0.1, 0.1, 0.7 |
| 3 | $-2$ | 0.8, 0.2 | 0.8, 0.2 | 0.8, 0.1, 0.1 | 0.8, 0.1, 0.1 | 0.25, 0.25, 0.25, 0.25 |

We can then compute the predicted hazard ratios for each iteration $r$ of the MCMC as

$$\frac{\exp(\hat{\theta}_s^r)}{\exp(\hat{\theta}_1^r)} \tag{31}$$

where $\hat{\theta}_1$ is the baseline hazard function, for example chosen as corresponding to the lowest values of all risk factors.

We can also compute the posterior predictive distribution of survival time as the expectation of the Weibull distribution, which is given by

$$\hat{y}_{pred} = \min(\hat{\gamma}_{c,r}^{-1/\bar{v}_c} \Gamma(1 + 1/\bar{v}_c), T^*)$$

where $T^*$ is the maximum observed survival time before censoring, $\bar{v}$ is the posterior mean of $v_c$ and $\hat{\gamma}^{c,r} = \exp(\hat{\theta}_s^r + \hat{\boldsymbol{\beta}}^T \mathbf{w})$ with $\hat{\boldsymbol{\beta}}$ the posterior mean of $\boldsymbol{\beta}$.

## 6 Application to simulated data

First we demonstrate the proposed method on two simulated datasets and then compare the results to ridge regression, a commonly used method to deal with collinearity.

### 6.1 Clustering censored survival data

We demonstrate the proposed method on two datasets, each with three clusters of 250 observations. We simulated the response variables $y$ from a Weibull distribution and five covariates $x_1, x_2,\ldots, x_5$ are drawn from multinomial distributions with 2, 2, 3, 3, and 4 categories respectively. There are no fixed effects. The values of the parameters $\theta$ and $\phi$ for each cluster $c$ are given in Table 1. The shape parameter for the first dataset is $v_c = 2$ for all clusters $c = 1, 2, 3$. The second dataset has cluster-specific shape parameters, so $v_1 = 2$, $v_2 = 1$ and $v_3 = 3$. There are no missing values but all observations are censored at $y = 50$ if the event has not happened yet.

We analyse the datasets using the proposed survival profile regression, implemented in the R package PReMiuM (Liverani et al. 2015) with 2000 iterations of burn in period and 2000 iterations after burn in. Good convergence (and mixing) of MCMC output was achieved within a few hundred iterations (based on visual diagnosis of MCMC output for model parameters, not shown).
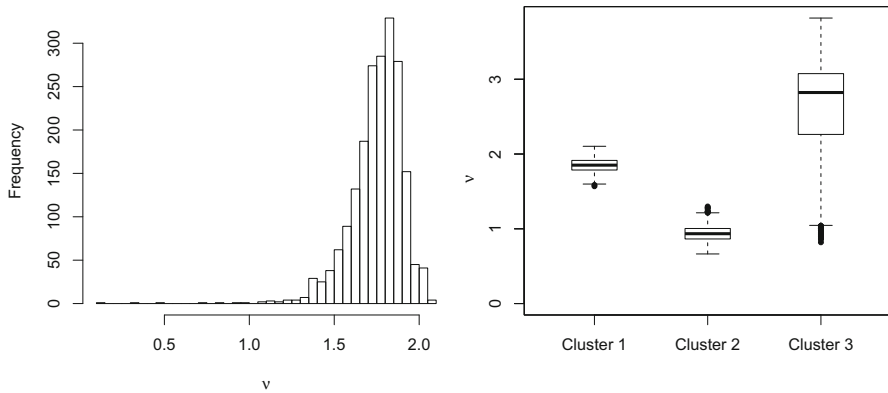
**Fig. 1** The posterior distribution of $\nu$ for the first dataset (left hand side) and $\boldsymbol{\nu}$ for the second dataset (right hand side)
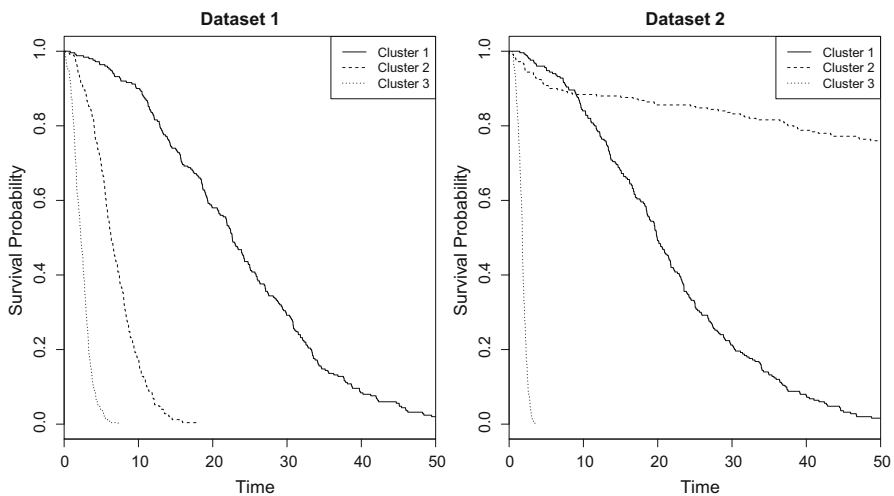


**Fig. 2** Survival probability for the two simulated datasets. Each survival function corresponds to a different cluster

The first dataset was analysed using the model with global shape parameter, while the second was analysed using a cluster-specific shape parameter. The posterior distributions are consistent with the generating values provided in Table 1 and they are shown in the Appendix. We show here the posterior distribution of $\nu$ for the first dataset and the posterior distribution of $\boldsymbol{\nu}$ for the second dataset in Fig. 1. The survival probability over time for the three clusters is given in Fig. 2.

As discussed, we could compute hazard ratios for the first dataset as we have assumed a global shape parameter. However, for the second dataset, generated with cluster-specific parameters, the analysis has allowed the shape parameter to be cluster specific and found that it was different for the three clusters. The log cumulative hazard function, shown in Fig. 3, shows that the assumption of proportional hazards
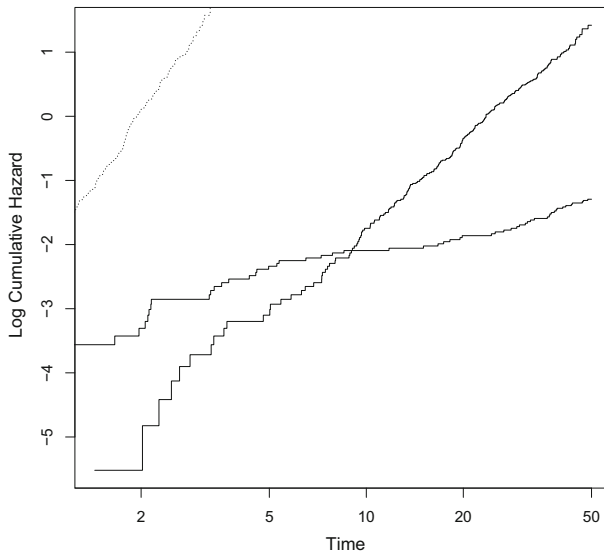
**Fig. 3** Log cumulative hazard for the three clusters of the second dataset

does not hold, and therefore we cannot compute hazard ratios meaningfully. Instead, for example, we can compare the clusters and interpret the results using the posterior survival time.

## 6.2 Comparison with ridge regression

We compare our clustering method to ridge regression (Gray 1992; Xue et al. 2007), a method suitable for collinear survival data. We generated 50 datasets with three 2-dimensional clusters, where the two variables are highly correlated within each cluster. The three clusters, of 300, 400 and 500 observations each, are generated from a bivariate Normal distribution with correlation of 0.95. The survival time is also generated from a Normal distribution. A censoring variable is generated from a Binomial distribution with $p = 0.9$, so only about 10

As a measure of accuracy for profile regression and ridge regression, we compare their predictive power using the root mean square error (RMSE) of the predicted values with respect to the observed outcome. This measure of goodness of fit is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left( y_i - \hat{y}_i \right)^2}{n}} \tag{32}$$

where $\hat{y}_i$ denotes the mean of the posterior predictive distribution for the survival time for observation $i$. Table 2 shows the mean and the standard deviation of the RMSE for the 50 simulated datasets. The precision of the in-sample predicted survival times obtained with profile regression was higher than the one obtained with ridge regression.

**Table 2** Mean and standard deviation of the RMSE of the predicted values with respect to the observed outcome when using profile regression and ridge regression

|  | Mean (RMSE) | SD (RMSE) |
|---|---|---|
| Profile regression | 1.91 | 0.09 |
| Ridge regression | 7.43 | 0.01 |

## 7 Application to the ALSWH sleep data

We apply our two methods to The Australian Longitudinal Study on Women's Health (ALSWH), a longitudinal study of over 40,000 women, consisting of three cohorts. The women were randomly selected from the Australian national health insurance database (Medicare), with oversampling of women from rural and remote areas to allow adequate numbers for statistical comparisons to be made. At baseline, in 1996, the cohorts, known according to the year the women were born as '1973–78', '1946–51', and '1921–26', were aged 18–23, 45–50, and 70–75. Follow-up omnibus style surveys were mailed out every three years. The ALSWH explores factors that influence health among women who are broadly representative of the entire Australian population, and is the largest project of its kind ever conducted in Australia. The current analysis focuses on data from the oldest cohort, born between 1921–26, who completed the baseline survey in 1996, and who first completed the sleep questionnaire 3 years later at Survey 2 (N = 10076).

We carried out the analysis of the data using the proposed survival profile regression. The response variable of interest is survival, measured in years from Survey 2. Deaths were ascertained from the National Death Index (Powers et al. 2000). The data cover 16 years after the survey and there is significant censoring: 97 women were not followed up in any survey and 5144 were still alive at the last survey. Sleep difficulty was measured using items from the NHP (Nottingham Health Profile) Sleep subscale (Hunt et al. 1981), as follows:

1. Do you wake in the early hours of the morning? (*early*)
2. Do you lie awake most of the night? (*lying*)
3. Do you take a long time to get to sleep? (*long*)
4. Do you sleep badly at night? (*bad*)

We will refer to these sleep items using the words provided in the parenthesis next to each item. The answers to these questions were coded as *no* = 0 and *yes* = 1. We also surveyed use of sleep medication (*meds*), first measured at Survey 2 (referred to as 'baseline' sleep difficulty). Other covariates of interest, measured at baseline, included comorbidity count (*comorb*, classified as 0, 1–2, and 3 or more[1]), marital status (*ms*, classified as married/de facto, separated/divorced, widowed, never married), area of residence (*area*, classified as Major Cities of Australia, Inner Regional Australia, Outer Regional Australia, Remote/Very Remote Australia) (Department

[1] Women were questioned about diagnosed medical conditions, including diabetes, arthritis, heart disease, hypertension, asthma, bronchitis/emphysema, stroke, osteoporosis and cancer. The total number of reported diseases at baseline was categorised as none, 1–2, or 3 or more. These three categories were utilised to reflect the varying severity of disease and comorbid conditions (no disease, disease with no or only a single comorbid condition, and multiple comorbid conditions).
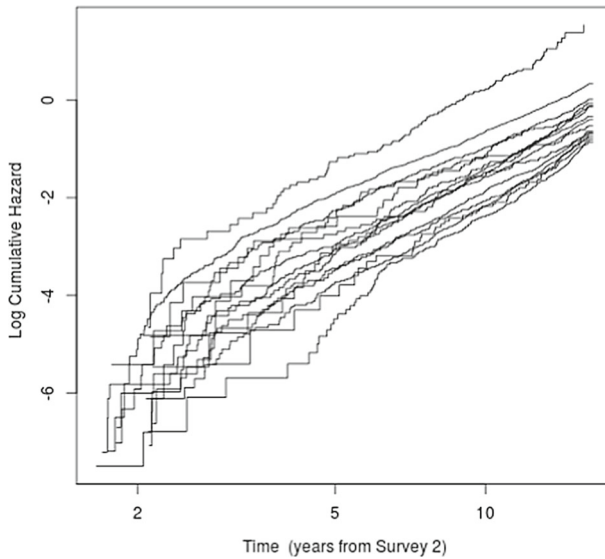
**Fig. 4** Log cumulative hazard function for the clusters

of Health and Aged Care 2001), highest obtained educational status (*edu*, classified as none, school/higher school certificate, trade/diploma, higher education), self-rated health (*srgood*, classified as excellent/very good/good or fair/poor), Short Form Health Survey (SF36) (Ware et al. 1994) measures of physical functioning (*pfq*), mental health (*mhq*) and vitality (*vtq*, classified on its quartiles), and the body mass index (*bmi*, classified as underweight, normal weight, overweight or obese). Age (years) at baseline was also included as a fixed effect. Due to the fact that the survival profile regression and sleep/disease profiles are estimated simultaneously in the current analysis, we restricted the profiles to baseline data only (as opposed to longitudinal), to avoid the situation where missing data due to death at later surveys might dominate the resultant profiles. However, prior work (Leigh et al. 2015) investigating longitudinal patterns of sleep difficulty has shown that sleep difficulty patterns remain stable over time, and thus the baseline values are fairly representative of the women's sleep patterns over time.

We obtained fifteen clusters. The credible interval for $\beta$ is (0.14,0.18). Figure 4 shows that the log cumulative hazard function does not support the assumption of proportional hazards and thus a cluster-specific shape parameter was used in the modelling. Therefore we do not compute hazard ratios but analyse the data using survival times. The cluster sizes and the posterior means of the parameters $\theta$ and $\phi$ are given in Table 3, with the clusters ordered according to their estimated mean survival time. Figures 5 and 6 show the posterior survival probabilities for the clusters. It can be seen that the survival functions for the first five clusters are distinct, while they are clustered together and overlapping for the remaining clusters.

**Table 3** Cluster sizes and posterior means for the cluster specific parameters $\theta$ and $\phi$

| Cluster | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size | | 106 | 1332 | 455 | 752 | 236 | 128 | 1356 | 1125 | 1182 | 223 | 887 | 1811 | 124 | 816 | 227 |
| $\theta_c$ | | 13.99 | 15.31 | 16.61 | 17.15 | 18.54 | 18.83 | 19.26 | 19.55 | 20.1 | 20.19 | 20.65 | 20.74 | 20.78 | 21.57 | 22.05 |
| Early | 0 | 0.64 | 0.56 | 0.31 | 0.54 | 0.44 | 0.53 | 0.56 | 0.67 | 0.49 | 0.27 | 0.53 | 0.73 | 0.66 | 0.59 | 0.4 |
| | 1 | 0.36 | 0.44 | 0.69 | 0.46 | 0.56 | 0.47 | 0.44 | 0.33 | 0.51 | 0.73 | 0.47 | 0.27 | 0.34 | 0.41 | 0.6 |
| Lying | 0 | 0.92 | 0.86 | 0.36 | 0.91 | 0.74 | 0.84 | 0.96 | 0.97 | 0.86 | 0.36 | 0.94 | 0.98 | 0.85 | 0.94 | 0.48 |
| | 1 | 0.08 | 0.14 | 0.64 | 0.09 | 0.26 | 0.16 | 0.04 | 0.03 | 0.14 | 0.64 | 0.06 | 0.02 | 0.15 | 0.06 | 0.52 |
| Long | 0 | 0.8 | 0.68 | 0.28 | 0.79 | 0.48 | 0.67 | 0.84 | 0.84 | 0.67 | 0.24 | 0.78 | 0.88 | 0.38 | 0.79 | 0.41 |
| | 1 | 0.2 | 0.32 | 0.72 | 0.21 | 0.52 | 0.33 | 0.16 | 0.16 | 0.33 | 0.76 | 0.22 | 0.12 | 0.62 | 0.21 | 0.59 |
| Bad | 0 | 0.86 | 0.73 | 0.26 | 0.82 | 0.54 | 0.6 | 0.87 | 0.91 | 0.71 | 0.22 | 0.81 | 0.93 | 0.77 | 0.86 | 0.42 |
| | 1 | 0.14 | 0.27 | 0.74 | 0.18 | 0.46 | 0.4 | 0.13 | 0.09 | 0.29 | 0.78 | 0.19 | 0.07 | 0.23 | 0.14 | 0.58 |
| Comorb | 0 | 0.15 | 0.05 | 0.06 | 0.06 | 0.08 | 0.24 | 0.16 | 0.18 | 0.12 | 0.13 | 0.24 | 0.37 | 0.17 | 0.3 | 0.2 |
| | 1 | 0.49 | 0.43 | 0.4 | 0.54 | 0.45 | 0.56 | 0.62 | 0.65 | 0.62 | 0.6 | 0.63 | 0.56 | 0.54 | 0.6 | 0.58 |
| | 2 | 0.36 | 0.52 | 0.54 | 0.4 | 0.47 | 0.18 | 0.22 | 0.17 | 0.26 | 0.27 | 0.13 | 0.07 | 0.28 | 0.1 | 0.2 |
| ms | 1 | 0.23 | 0.22 | 0.22 | 0.23 | 0.23 | 0.22 | 0.23 | 0.23 | 0.22 | 0.23 | 0.23 | 0.22 | 0.23 | 0.23 | 0.21 |
| | 2 | 0.24 | 0.22 | 0.22 | 0.22 | 0.23 | 0.24 | 0.23 | 0.23 | 0.22 | 0.22 | 0.23 | 0.22 | 0.23 | 0.23 | 0.22 |
| | 3 | 0.23 | 0.22 | 0.21 | 0.23 | 0.23 | 0.22 | 0.22 | 0.24 | 0.22 | 0.22 | 0.23 | 0.22 | 0.24 | 0.23 | 0.23 |
| | 4 | 0.24 | 0.22 | 0.23 | 0.24 | 0.23 | 0.23 | 0.23 | 0.23 | 0.23 | 0.23 | 0.23 | 0.22 | 0.23 | 0.23 | 0.23 |
| Area | 1 | 0.23 | 0.22 | 0.22 | 0.22 | 0.23 | 0.23 | 0.23 | 0.23 | 0.23 | 0.22 | 0.23 | 0.23 | 0.24 | 0.23 | 0.22 |
| | 2 | 0.23 | 0.21 | 0.22 | 0.23 | 0.23 | 0.23 | 0.23 | 0.23 | 0.22 | 0.23 | 0.23 | 0.22 | 0.25 | 0.23 | 0.23 |
| | 3 | 0.24 | 0.22 | 0.23 | 0.23 | 0.23 | 0.22 | 0.23 | 0.22 | 0.23 | 0.23 | 0.23 | 0.22 | 0.23 | 0.22 | 0.23 |
| | 4 | 0.24 | 0.22 | 0.22 | 0.23 | 0.23 | 0.22 | 0.24 | 0.23 | 0.22 | 0.21 | 0.23 | 0.22 | 0.23 | 0.23 | 0.21 |
| Edu | 1 | 0.26 | 0.39 | 0.45 | 0.27 | 0.25 | 0.23 | 0.23 | 0.22 | 0.24 | 0.25 | 0.22 | 0.24 | 0.23 | 0.22 | 0.23 |
| | 2 | 0.27 | 0.42 | 0.39 | 0.27 | 0.25 | 0.27 | 0.26 | 0.3 | 0.26 | 0.25 | 0.45 | 0.5 | 0.27 | 0.28 | 0.26 |
| | 3 | 0.21 | 0.12 | 0.1 | 0.2 | 0.22 | 0.21 | 0.22 | 0.21 | 0.21 | 0.21 | 0.19 | 0.16 | 0.22 | 0.21 | 0.21 |

**Table 3** continued

| Cluster | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| srgood | 4 | 0.2 | 0.06 | 0.05 | 0.18 | 0.21 | 0.21 | 0.21 | 0.18 | 0.2 | 0.2 | 0.13 | 0.09 | 0.21 | 0.19 | 0.2 |
| | 1 | 0.51 | 0.78 | 0.77 | 0.43 | 0.33 | 0.1 | 0.15 | 0.05 | 0.34 | 0.26 | 0.03 | 0.01 | 0.16 | 0.07 | 0.06 |
| | 2 | 0.49 | 0.22 | 0.23 | 0.57 | 0.67 | 0.9 | 0.85 | 0.95 | 0.66 | 0.74 | 0.97 | 0.99 | 0.84 | 0.93 | 0.94 |
| Meds | 1 | 0.82 | 0.73 | 0.57 | 0.85 | 0.65 | 0.47 | 0.91 | 0.92 | 0.76 | 0.68 | 0.84 | 0.94 | 0.56 | 0.9 | 0.75 |
| | 2 | 0.18 | 0.27 | 0.43 | 0.15 | 0.35 | 0.53 | 0.09 | 0.08 | 0.24 | 0.32 | 0.16 | 0.06 | 0.44 | 0.1 | 0.25 |
| pfq | 1 | 0.3 | 0.72 | 0.66 | 0.57 | 0.21 | 0.09 | 0.15 | 0.1 | 0.27 | 0.26 | 0.04 | 0.02 | 0.11 | 0.06 | 0.08 |
| | 2 | 0.32 | 0.2 | 0.22 | 0.27 | 0.43 | 0.17 | 0.33 | 0.24 | 0.36 | 0.35 | 0.13 | 0.06 | 0.37 | 0.16 | 0.18 |
| | 3 | 0.23 | 0.07 | 0.08 | 0.12 | 0.28 | 0.21 | 0.36 | 0.36 | 0.25 | 0.24 | 0.34 | 0.22 | 0.36 | 0.3 | 0.34 |
| | 4 | 0.15 | 0.02 | 0.04 | 0.03 | 0.07 | 0.53 | 0.16 | 0.31 | 0.12 | 0.14 | 0.49 | 0.69 | 0.15 | 0.48 | 0.41 |
| mhq | 1 | 0.19 | 0.5 | 0.63 | 0.16 | 0.16 | 0.37 | 0.13 | 0.05 | 0.49 | 0.39 | 0.07 | 0.03 | 0.19 | 0.24 | 0.09 |
| | 2 | 0.29 | 0.24 | 0.21 | 0.24 | 0.26 | 0.34 | 0.25 | 0.1 | 0.28 | 0.28 | 0.18 | 0.08 | 0.2 | 0.31 | 0.15 |
| | 3 | 0.29 | 0.18 | 0.12 | 0.39 | 0.43 | 0.17 | 0.42 | 0.31 | 0.17 | 0.24 | 0.45 | 0.3 | 0.4 | 0.35 | 0.39 |
| | 4 | 0.23 | 0.07 | 0.04 | 0.22 | 0.15 | 0.12 | 0.21 | 0.54 | 0.05 | 0.09 | 0.29 | 0.59 | 0.21 | 0.11 | 0.37 |
| vtq | 1 | 0.27 | 0.71 | 0.71 | 0.33 | 0.22 | 0.1 | 0.12 | 0.03 | 0.5 | 0.34 | 0.03 | 0.01 | 0.14 | 0.09 | 0.05 |
| | 2 | 0.37 | 0.23 | 0.23 | 0.38 | 0.43 | 0.37 | 0.38 | 0.13 | 0.36 | 0.39 | 0.17 | 0.05 | 0.35 | 0.36 | 0.15 |
| | 3 | 0.18 | 0.04 | 0.04 | 0.19 | 0.22 | 0.28 | 0.31 | 0.23 | 0.1 | 0.16 | 0.29 | 0.12 | 0.25 | 0.33 | 0.26 |
| | 4 | 0.17 | 0.02 | 0.01 | 0.1 | 0.12 | 0.25 | 0.19 | 0.61 | 0.04 | 0.1 | 0.51 | 0.82 | 0.24 | 0.22 | 0.54 |
| bmi | 1 | 0.2 | 0.06 | 0.04 | 0.06 | 0.16 | 0.07 | 0.14 | 0.04 | 0.17 | 0.11 | 0.06 | 0.04 | 0.16 | 0.06 | 0.11 |
| | 2 | 0.49 | 0.34 | 0.35 | 0.27 | 0.28 | 0.63 | 0.33 | 0.38 | 0.28 | 0.36 | 0.61 | 0.61 | 0.32 | 0.59 | 0.37 |
| | 3 | 0.18 | 0.34 | 0.38 | 0.35 | 0.27 | 0.24 | 0.26 | 0.38 | 0.23 | 0.34 | 0.26 | 0.29 | 0.26 | 0.26 | 0.3 |
| | 4 | 0.12 | 0.26 | 0.23 | 0.31 | 0.21 | 0.06 | 0.18 | 0.17 | 0.2 | 0.16 | 0.07 | 0.06 | 0.19 | 0.07 | 0.15 |
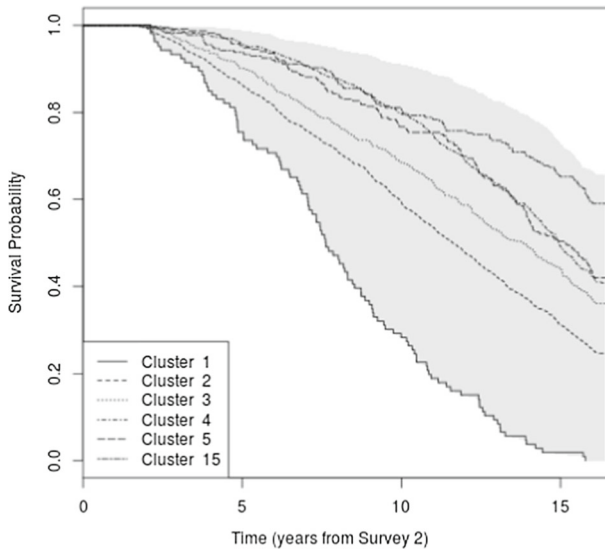
**Fig. 5** Survival probability for the first 5 clusters and the last cluster. The gray area highlights the span of the survival probability for all clusters
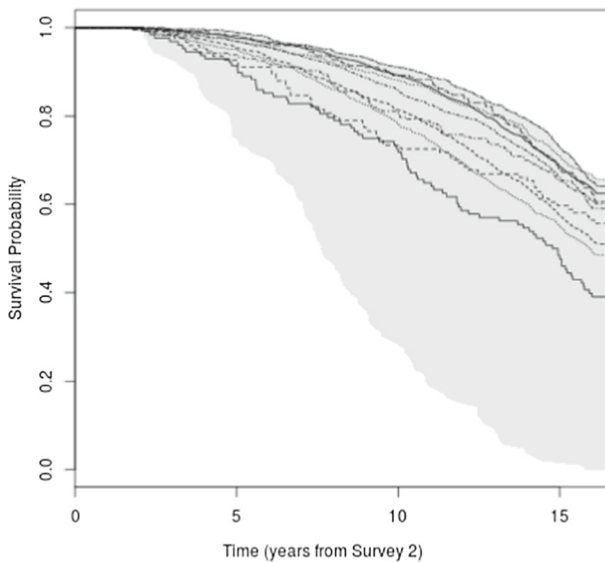


**Fig. 6** Survival probability for all clusters except the first five. The gray area highlights the span of the survival probability for all clusters

Figure 7 shows the boxplots for the posterior survival time for the fifteen clusters. The overall median is also shown in the plot, allowing a comparison with the deviation from the median of the posterior survival time for each cluster.
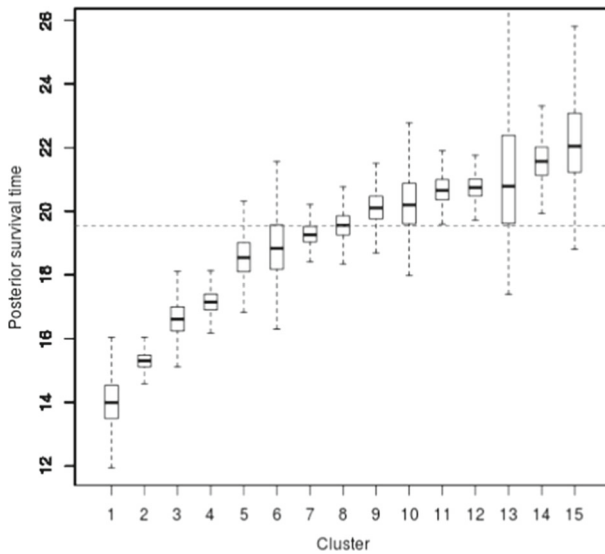
**Fig. 7** Posterior distribution of the survival time since Survey 2 for each cluster. The dashed horizontal line is the overall median survival time

We also carried out variable selection. Values of $\rho$ close to 1 indicate the variable is significant for the clustering, while values close to zero indicate it is not. The posterior distribution for $\rho$ showed that two of the covariates, *ms* and *area*, were not relevant for the clustering model, since the posterior distribution of $\rho$ for these two variables was heavily distributed close to zero. This is demonstrated in Fig. 8, which shows the posterior distribution of $\rho$ for marital status and area, as well as example distributions of $\rho$ for variables which are important for the model (*early* and *lying*). The distribution of $\rho$ for these latter two variables is distributed closer to 1.

We propose the use of a heatmap as the most immediate way to visualise the clustering and associated covariate patterns. Figure 9 shows a summary table of the survival time and the posterior distribution of the covariates in each cluster. Each row represents a cluster. The columns represents, respectively, the mean survival time and each covariate included in the analysis. The colour of each cell in the matrix corresponds to a quintile of the distribution for that variable (ie. by column). The clusters are ordered as in Table 3 by survival time, from the shortest to the longest. Note that the colours in the matrix do not become darker (or lighter) in a smooth manner, suggesting a complex relationship between survival time and the covariates considered. For example, we can see that higher levels of physical functioning and vitality are generally associated with longer average survival times. However, there are several exceptions to this, and we can see complex non-linear relationships between survival time and the other covariates (Table 4).

We then exclude the covariates which are not driving the clustering process, as identified by looking at the posterior distribution of $\rho$. Each value in the heatmap gives the quintiles of the distribution, therefore summarising the clusters. It can be noted how the relationships between the covariates are complex and could not have been easily
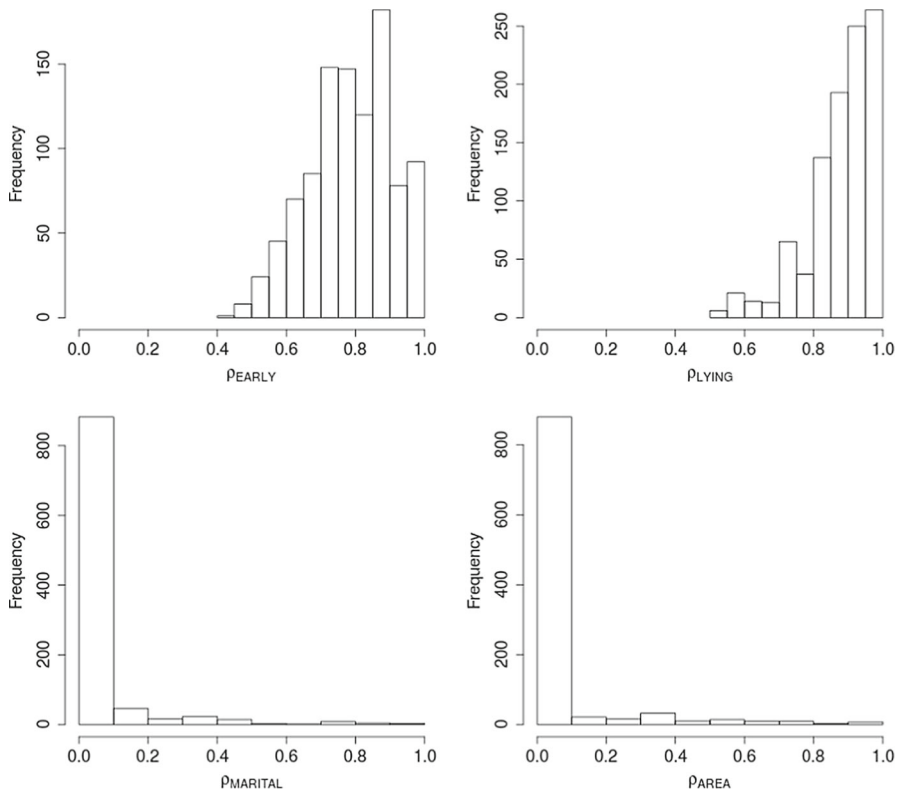
**Fig. 8** Posterior distribution of the parameter $\rho$ for four covariates: *early*, *lying*, *ms* and *area*

learnt using other methods. We can see that there are three clusters (3, 10 and 15) with overall poor sleep difficulty patterns. Of these clusters, two (10 and 15) correspond to long median survival times, while one of them (3) corresponds to a shorter median survival time. Three other clusters had individuals with greater sleep difficulty patterns in just some of the domains (clusters 5, 9 and 13). Moreover, we can see how these clusters are also associated with other covariates such as medication, high BMI or low levels of vitality and physical activity. The cluster with the lowest survival had a high probability of comorbidities, but reported no sleep difficulty. Cluster 3, which reported sleep difficulty and the 3rd shortest median survival, did not exhibit high probability of comorbidities, however they scored low on all QoL covariates, and were likely to have higher BMIs. The two other clusters with greatest sleep difficulty, 10 and 15, exhibited good self-rated health, low to moderate likelihood of using sleep meds, cluster 10 did not score high on all quality of life (QoL) items but cluster 15 did, and neither exhibited high BMIs. Cluster 13 endorsed 'taking a long time to get to sleep', and also were likely to use meds, have high QoL, and good self-rated health. Cluster 9 endorsed 'early waking', had good self-rated health but moderate QoL measures. Many clusters were characterised by low probability of sleep difficulty across all items.
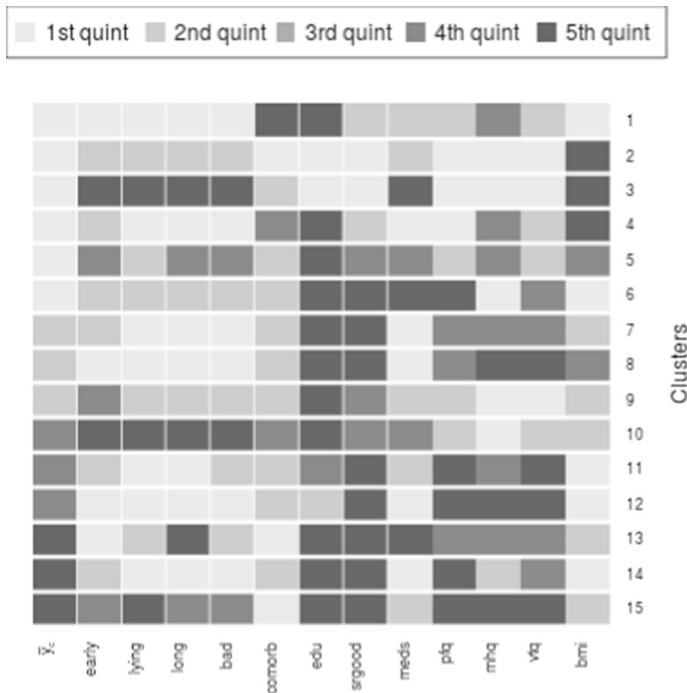
**Fig. 9** Heatmap summary table of the clusters. Each row represents a cluster. The columns represents, respectively, the mean survival time and each covariate included in the analysis. The colour of each cell in the matrix corresponds to a quintile of the distribution for that variable (ie. by column). The clusters are ordered as in Table 3 by survival time, from the shortest to the longest

We can thus learn and visualise how the posterior distribution of median survival time changes depending on the values of different covariates. For example, Fig. 10 shows the posterior predictive distributions for three profiles of women who answer yes to the question 'Do you wake in the early hours of the morning?' (*early = 1*). For the first profile the individual is healthy and their sleep patterns are good based on their answers (*no*) to the other items of the Nottingham Health Profile (*lying = 0, long = 0, bad = 0*). For the second profile, they did not answer the other sleep questions (*lying = NA, long = NA, bad = NA*) but they are healthy. For the third profile, we only know that they wake up early. The posterior predictive distributions of these three profiles allow us to make inference on the median survival times for specific individuals, or groups of individuals, and shed light on the potentially complex true relationsips between covariates and survival, as highlighted by the multimodality of the posterior predictive distributions shown in Fig. 10.

**Table 4** Posterior mean of the probabilities of $\phi$ for the first simulated dataset

| Covariate | Category | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|---|
| $x_1$ | 0 | 0.73 | 0.77 | 0.26 |
|  | 1 | 0.27 | 0.23 | 0.74 |
| $x_2$ | 0 | 0.22 | 0.78 | 0.20 |
|  | 1 | 0.78 | 0.22 | 0.80 |
| $x_3$ | 0 | 0.12 | 0.79 | 0.15 |
|  | 1 | 0.09 | 0.08 | 0.09 |
|  | 2 | 0.78 | 0.13 | 0.76 |
| $x_4$ | 0 | 0.12 | 0.78 | 0.10 |
|  | 1 | 0.81 | 0.10 | 0.82 |
|  | 2 | 0.07 | 0.12 | 0.08 |
| $x_5$ | 0 | 0.21 | 0.20 | 0.14 |
|  | 1 | 0.27 | 0.24 | 0.09 |
|  | 2 | 0.22 | 0.29 | 0.14 |
|  | 3 | 0.31 | 0.27 | 0.63 |
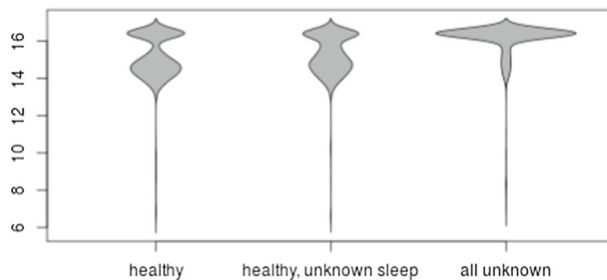


**Fig. 10** Posterior predictive distributions for three profiles of interest. The three profiles show the posterior distribution for predictive profiles for individuals who replied *yes* to the question 'Do you wake in the early hours of the morning?'. For the first profile on the left hand side we also know that the individual is healthy and their sleep patterns are good otherwise (*lying = no, long = no, bad = no*). For the second profile, we have no knowledge of how they answered the other sleep questions (*lying = NA, long = NA, bad = NA*) but know that they are healthy. For the third profile, we have no knowledge about the individual apart from the fact that they wake up early

## 8 Discussion

We have proposed a mixture model for the survival response and covariates, where the response variable has a Weibull distribution and it allows for censoring. In the model we allow for the shape parameter of the Weibull distribution to be shared by all clusters (therefore satisfying the condition of proportional hazards) but also proposed a more general model with cluster-specific shape parameters. Moreover, we have discussed the challenges of predictive profiles in the context of survival modelling and we have made these methods easily available in the R package PReMiuM.

We have used the latter model to analyse data from The Australian Longitudinal Study on Women's Health and have demonstrated how useful inference can be drawn
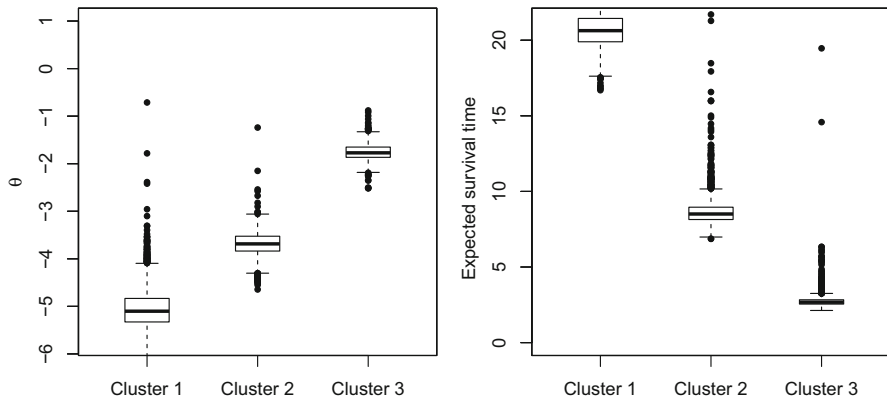
**Fig. 11** The posterior distribution of $\theta$ and the posterior distribution of the survival time for the first simulated dataset
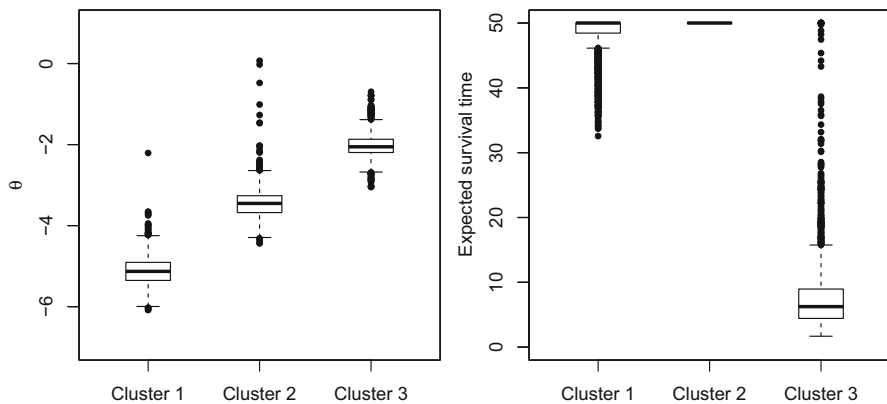


**Fig. 12** Posterior distribution of $\theta$ and the posterior distribution of the survival time for the second simulated dataset

using our proposed models. A previous analysis (Leigh et al. 2015), which clustered the women based only on the sleep difficulty questions, found four clusters, corresponding to no sleep difficulty (answered *no* to all questions on sleep), trouble sleeping (answered *yes* to all questions on sleep), early wakers, and trouble falling asleep. The current clustering also identified clusters characterised by low sleep difficulty, trouble sleeping, and a cluster defined by taking a long time to get to sleep (those who answered *yes* to whether they take a long time to get to sleep). In the current analysis, many more clusters were identified as additional covariate data was also allowed to inform the clustering (Figs. 11, 12).

Leigh et al. (2016a, b) found that, unadjusted for covariates, those with mild sleep difficulty had lower hazard of death than those without sleep difficulty, while the most troubled sleepers had higher hazard of death. After adjusting for covariates, the troubled sleepers did not have greater hazard of death, and those with mild sleep difficulty (early wakers and trouble falling asleep) still had lower hazard of death. Also

significant in the models were disease count, BMI, education, physical functioning, self-rated health, marital status and area. The effect of those covariates, in conjunction with sleep, led to many more clusters being identified in the current analysis. We observe that greater sleep difficulty can be related to both longer and shorter survival, with different patterns in the covariates. This may be attributable to the difference in the effect of trouble sleeping with and without covariate adjustment in the previous models. Furthermore, a previous analysis (Leigh et al. 2015) could not account for the interaction between sleep and each covariate of interest. The current analysis however takes into account the multivariate relationships between all variables, and leads to interesting insights. For example, previous work Leigh et al. (2016b) found that BMI was significant for survival when modelling as a predictor along side sleep, with underweight related to greater hazard of death, overweight greater hazard of death, and non significant results for obese women. However, BMI may also be interrelated with sleep, for instance obesity is related to sleep apnoea, which can cause sleep disturbance. While we see one class with high BMI and poor sleep and shorter survival (cluster 3), the other clusters with the greatest sleep difficulty do not have higher BMIs, and also have longer survival. It is possible that the relationship between greater sleep difficulty and survival in cluster 3 is explained by the high BMI, whereas the relationship between sleep difficulty and lower hazard of death in clusters 10 and 15 is explained by healthier BMI (and better self-rated health etc.).

Moreover, in previous work clustering was conducted prior to regression on the survival outcome, and thus missing covariate data patterns related to survival may have influenced class membership. Thus, the clusters themselves may have also included information about the survival outcome, and thus possibly biased the subsequent regression results. The current analysis used only baseline data, thus the clusters themselves are not dominated by missing data.

A limitation of our model is that in its present formulation does not incorporate covariate information collected after the baseline, which is the objective for our future work.

**Data Availability Statement** The code used to simulate data and analyse them is included in the supplementary information files of this paper. The data owned by the ALSWH can be accessed with permission from the Data Access Committee of ALSWH.
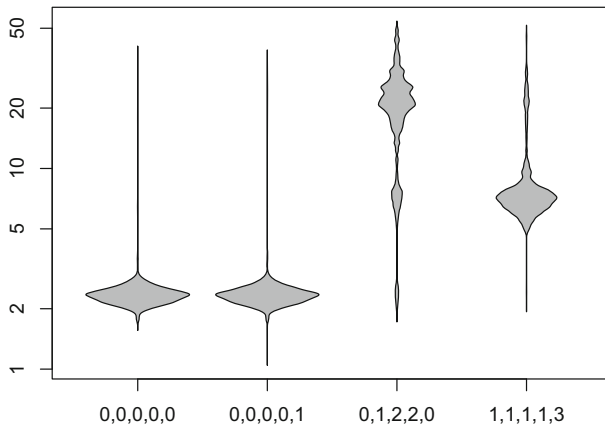
**Fig. 13** Violin plot of the posterior predictive distribution for four predictive profiles: (0,0,0,0,0), (0,0,0,0,1), (0,1,2,2,0) and (1,1,1,1,3) for the second simulated dataset

## Appendix: Additional results from the simulation study

In this Appendix we provide additional plots showing the posterior distribution for some of the parameters for the simulated data in Sect. 6.

We also show in Fig. 13 the posterior predictive distribution for four predictive profiles. These are combinations of values of the covariates for which we compute the posterior predictive distribution of $\theta_c$ or the survival time.

## References

Bigelow JL, Dunson DB (2009) Bayesian semiparametric joint models for functional predictors. J Am Stat Assoc 104(485):26–36

Borzadaran GRM, Borzadaran HAM (2011) Log-concavity property for some well-known distributions. Surv Math Appl 6:203–219

Chung Y, Dunson DB (2009) Nonparametric bayes conditional distribution modeling with variable selection. J Am Stat Assoc 104(488):1646–1660

Coker E, Liverani S, Ghosh JK, Jerrett M, Beckerman B, Li A, Ritz B, Molitor J (2016) Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County. Environ Int 91:1–13

Department of Health and Aged Care (2001) Measuring remoteness: accessibility/remoteness Index of Australia (ARIA) revised edition, Volume 14. Occasional papers: new series

Dunson DB, Herring AB, Siega-Riz AM (2008) Bayesian inference on changes in response densities over predictor clusters. J Am Stat Assoc 103(484):1508–1517

Gilks WR, Wild P (1992) Adaptive rejection sampling for gibbs sampling. Appl Stat 41:337–348

Gray RJ (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. J Am Stat Assoc 87(420):942–951

Hastie DI, Liverani S, Azizi L, Richardson S, Stücker I (2013) A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer. BMC Med Res Methodol 13(1):129

Hastie DI, Liverani S, Richardson S (2015) Sampling from dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. Stat Comput 25(5):1023–1037

Hunt SM, McKenna SP, McEwen J, Williams J, Papp E (1981) The nottingham health profile: subjective health status and medical consultations. Soc Sci Med Part A Med Psychol Med Sociol 15(3):221–229

Kottas A (2006) Nonparametric Bayesian survival analysis using mixtures of weibull distributions. J Stat Plan Inference 136(3):578–596

Leigh L, Hudson IL, Byles JE (2015) Sleeping difficulty, disease and mortality in older women: a latent class analysis and distal survival analysis. J Sleep Res 24(6):648–657

Leigh L, Hudson IL, Byles JE (2016a) Joint modelling of the relationship between sleep, disease and mortality, exclusively in a cohort of older australian women (aged 70–75 years at baseline). J Stat Adv Theory Appl 16(2):185–254

Leigh L, Hudson IL, Byles JE (2016b) Sleep difficulty and disease in a cohort of very old women. J Aging Health 28(6):1090–1104

Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S (2015) PReMiuM: an R package for profile regression mixture models using dirichlet processes. J Stat Softw 64(7):1–30

Liverani S, Lavigne A, Blangiardo M (2016) Modelling collinear and spatially correlated data. Spatial Spatio-temporal Epidemiol 18:63–73

Mattei F, Liverani S, Guida F, Matrat M, Cenée S, Azizi L, Menvielle G, Sanchez M, Pilorget C, Lapôtre-Ledoux B et al (2016) Multidimensional analysis of the effect of occupational exposure to organic solvents on lung cancer risk: the ICARE study. Occup Environ Med 73(6):368–377

Molitor, J., I. J. Brown, Q. Chan, M. Papathomas, S. Liverani, N. Molitor, S. Richardson, L. Van Horn, M. L. Daviglus, A. Dyer, J. Stamler, P. Elliott, and I. R. Group (2014) Blood pressure differences associated with optimal macronutrient intake trial for heart health (OMNIHEART)-like diet compared with a typical American diet. Hypertension 64(6):1198–1204

Molitor J, Papathomas M, Jerrett M, Richardson S (2010) Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics 11(3):484–498

Molitor J, Su JG, Molitor N-T, Rubio VG, Richardson S, Hastie D, Morello-Frosch R, Jerrett M (2011) Identifying vulnerable populations through an examination of the association between multipollutant profiles and poverty. Environ Sci Technol 45(18):7754–7760

Papathomas M, Molitor J, Hoggart C, Hastie D, Richardson S (2012) Exploring data from genetic association studies using Bayesian variable selection and the dirichlet process: application to searching for gene× gene patterns. Genet Epidemiol 36(6):663–674

Papathomas M, Molitor J, Richardson S, Riboli E, Vineis P (2011) Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in nonsmokers. Environ Health Perspect 119(1):84

Pirani M, Best N, Blangiardo M, Liverani S, Atkinson RW, Fuller GW (2015) Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. Environ Int 79:56–64

Powers J, Ball J, Adamson L, Dobson A (2000) Effectiveness of the national death index for establishing the vital status of older women in the Australian longitudinal study on women's health. Aust N Z J Public Health 24(5):526–528

Teh YW (2011) Dirichlet process. In: Encyclopedia of machine learning, pp 280–287. Springer

Ware J, Kosinski M, Keller S (1994) Physical and mental health summary scales—a user's manual. New England Medical Center, The Health Institute, Boston

Xue X, Kim MY, Shore RE (2007) Cox regression analysis in presence of collinearity: an application to assessment of health risks associated with occupational radiation exposure. Lifetime Data Anal 13(3):333–350

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.