SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Original Research Article

Statistical Methods in Medical Research
2022, Vol. 31(5) 801–820

DOI: 10.1177/09622802211057290
journals.sagepub.com/home/smm
SAGE

Inference on latent factor models for informative censoring

Francesco Ungolo[1] and Edwin R. van den Heuvel[2]

Abstract

This work discusses the problem of informative censoring in survival studies. A joint model for the time to event and the time to censoring is presented. Their hazard functions include a latent factor in order to identify this joint model without sacrificing the flexibility of the parametric specification. Furthermore, a fully Bayesian formulation with a semi-parametric proportional hazard function is provided. Similar latent variable models have been described in literature, but here the emphasis is on the performance of the inferential task of the resulting mixture model with unknown number of components. The posterior distribution of the parameters is estimated using Hamiltonian Monte Carlo methods implemented in Stan. Simulation studies are provided to study its performance and the methodology is implemented for the analysis of the ACTG175 clinical trial dataset yielding a better fit. The results are also compared to the non-informative censoring case to show that ignoring informative censoring may lead to serious biases.

Keywords

Survival models, mixture models, bayesian inference, proportional hazard, HMC, model selection

Introduction

Right-censored survival times are very common in event time studies. Right-censoring is non-informative if the censoring times do not depend on the event of interest. For instance, units whose event of interest has not occurred by the end of the clinical study (type I censoring). Conversely, units may drop out from the study for reasons which depend on the event of interest. For example, in a clinical study on the efficacy of a new treatment, a patient may withdraw due to the worsening of their medical conditions.

In case of informative censoring, we need to consider a model for the joint distribution of the censoring time C and the time to event T. Not addressing this dependency between T and C may result in a biased estimation of the distribution of T. However, this joint distribution is not identifiable given the data, since we only observe the minimum between C and T (see[1],[2] and[3] for a more detailed account of this issue). Thus, we have to make untestable assumptions if we wish to study the joint distribution of T and C.

Some examples from the literature are the works of[4], who proposed and analysed the use of a bivariate Weibull model for (T, C), and the work of[5] who analyse the consistency of the estimator of the marginal survival function of T and C based on a copula model with known dependence parameters. These parametric models are typically used to investigate sensitivity with respect to the dependency parameter.

Scharfstein and Robins[6] consider a hazard function specification for the censoring time which has a multiplicative relationship with a function of T, while[7] specify a semi-parametric hazard function ([8]) for T, which has a step-change associated with the censoring event. In this way, they avoid modeling the marginal distribution of C. The limitation of these

[1] Chair of Mathematical Finance, Technical University of Munich, Garching bei München, Germany
[2] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

Corresponding author:
Francesco Ungolo, Chair of Mathematical Finance Technical University of Munich, Garching bei München, Germany.
Email: f.ungolo@tum.de

approaches is their need to fix the parameter which calibrates the degree of dependence between T and C, since it cannot be estimated on the basis of available data.

Huang and Wolfe[9] proposed the use of a latent variable to account for the dependency between T and C. Given this latent variable, then T and C are independently distributed. They assume that this latent variable is normally distributed with mean zero and unknown variance. This latent variable operates on a cluster level (e.g., clinics with patients) where each cluster has its own frailty, which is shared by all units in the cluster. The limitation of this approach is that the parametric class of distributions for the latent variable must be known.

Rowley et al.[10] extended the work of[9] by considering a latent factor with a finite but unknown number of possible outcomes. The resulting joint distribution of T and C turns out to be a mixture distribution with an unknown number of components. The authors suggested to estimate this joint distribution with the Maximum-a-Posteriori (MAP) estimator and they recommended to determine the number of levels with the Bayesian Information Criterion (BIC).

Among all options of modeling the joint distribution of T and C, the latent variable approach seems to be the most promising, since it has the least limitations compared to alternative parametric hazard function specifications.

However, statistical models including latent factors, such as those proposed in[10], contain singularities in their parameter space, which means that there is not a one-to-one mapping from the parameter space to a probability distribution. As a consequence, the Fisher Information Matrix may not be invertible (see[11] and[12]). Thus, estimators like maximum likelihood (ML) and MAP do not have asymptotic normal distributions and may yield divergent parameter estimates when applied to real datasets. The asymptotic distribution of the parameter estimators deviates from being Gaussian in ill-posed models, particularly when the estimator reaches the boundary of the parameter space ([13]). When dealing with mixture distributions this problem occurs when the mixture weight is close to zero, when two mixture components are very close to one another, or more generally when a mixture in K components can be equivalently obtained as a mixture in K' components, with $K' < K$.

The problem of model selection between distributions with different number of mixture components turns out to be extremely critical in this context. Classical information criteria such as the AIC ([14]) and the BIC ([15]) cannot be used (see[16] for further details).

Another criticism towards the use of ML and MAP estimators is that for a mixture model with K components the likelihood function is most likely multimodal. Fitting a K-component mixture model on n observations, results in a likelihood function of the form of a sum of K^n terms, each corresponding to the likelihood function obtainable under the possible realizations of the latent factor. Thus, there may exist multiple roots, and the observed solution can be affected by the choice of the starting values. Finally, the numerical optimization can require high computational time, due to the large number of parameters and the complexity of the objective function, and there would not be any guarantee that all roots will be found.

This work focuses on a joint model for T and C, where they are independently distributed conditional on a latent class or factor H, similar as in[10]. A point-identifying assumption for this distribution is specified, without sacrificing the flexibility of the parametric model.

We propose the use of a fully Bayesian analysis, where instead of focusing on a point estimate, we try to estimate the posterior distribution of the parameters. The contribution of this work is thus to address the inferential challenges posed by these mixture models for addressing informative censoring in survival models. We will discuss identifiability, the Bayesian inference for this model, including prior specification and parametrization of the mixture model, and the model selection problem for mixture distributions with unknown number of components.

Section 2 will discuss these model formulations, while Section 3 addresses the estimation problem within the Bayesian inferential paradigm. Section 4 contains a simulation study to assess the performance of the inferential strategy described in Section 3. As an illustrative example the ACTG175 clinical trial ([17]) is analysed, and the results are illustrated in Section 5. Extension of the use of joint models with latent factors are discussed in Section 6. Finally, Section 7 provides a summary and a discussion.

Modeling non-informative censoring with latent factors

In our model we assume that the event time T and the censoring time C are conditionally independent given latent factor H. Thus the joint density of (T, C) is given by:

$$f(t, c) = \int_h f_{T|H}\big(t \mid h\big) f_{C|H}\big(c \mid h\big) dF_H(h) \tag{1}$$

where $f(\cdot)$ denotes the density function while $F(\cdot)$ denotes the distribution function. We refer to the latent factor H as the frailty.

Furthermore, we allow that the distribution of T and C may depend on a vector of time varying covariates at time $u \in [0, \infty)$, denoted by $x^T(u) = (x_1^T(u), \ldots, x_{P_T}^T(u))$ and $x^C(u) = (x_1^C(u), \ldots, x_{P_C}^C(u))$ respectively. In this work, we include covariates by means of the Cox proportional hazard model ([8]). We assume that the distribution of H is independent of $x^T(\cdot)$ and $x^C(\cdot)$.

Model

We assume that the latent factor H is a discrete variable which takes $K+1$ possible values, denoted for simplicity by the set of integers $\{0, 1, \ldots, K\}$. K is considered to be fixed but unknown. We set $H = 0$ as the baseline level, and write $h = (h_1, \ldots, h_K)$, such that $h_k = 1_{[H=k]}$.

The following models for the hazard function of T and C are considered:

$$\mu^T\big(u \mid x^T(u), h; \beta^T, \gamma^T\big) = \mu_0^T(u) \exp\big(\beta^{T'} x^T(u) + \gamma^{T'} h\big) \tag{2}$$

$$\mu^C\big(u \mid x^C(u), h; \beta^C, \gamma^C\big) = \mu_0^C(u) \exp\big(\beta^{C'} x^C(u) + \gamma^{C'} h\big) \tag{3}$$

where β^ℓ is the vector of regression coefficients capturing the effect of the covariates $x^\ell(\cdot)$, and $\mu_0^\ell(\cdot)$ is the baseline hazard function. The unknown parameters $\gamma^\ell = (\gamma_0^\ell, \ldots, \gamma_K^\ell)$ capture the proportional effect of each outcome of H on the the hazard ($\ell \in \{T, C\}$). We assume that $\gamma_0^\ell = 0$.

As result, now the frailty component has a discrete distribution with sample space $(1, \exp(\gamma_1^\ell), \ldots, \exp(\gamma_K^\ell))$, for $\ell \in \{T, C\}$.

[10] also allow for the possibility of regression parameters β which depend on H, and the distribution of H may depend on X. An even more structured model, would allow H to be time-varying.

The latent factor H can be interpreted as an underlying health state or situation of a (group of) patient(s). For example, whether a patient used zidovudine for HIV Type I infection. In this case, the number of levels K for latent factor H is known and equal to two.

The resulting joint distribution of T and C still corresponds to the integral of equation (1), where the frailty component is dominated by the counting measure, hence the density can be rewritten as follows:

$$\begin{aligned} & f_{T,C|X^T,X^C}\big(t, c \mid x^T(t), x^C(c); \beta^T, \gamma^T, \beta^C, \gamma^C, \zeta\big) \\ & \quad = \sum_{k=0}^{K} \zeta_k f_{T|X^T,H}\big(t \mid x^T(t), h_k\big) f_{C|X^C,H}\big(c \mid x^C(c), h_k\big) \end{aligned} \tag{4}$$

where $\zeta = (\zeta_0, \ldots, \zeta_K)$ and $\zeta_k = \Pr(\mathrm{H} = \mathrm{k})$. Equation (4) is in fact a mixture distribution with $K+1$ components induced by the presence of a latent factor H. ζ can be considered as the non-parametric distribution of H (see[18,19,20]).

The mixture formulation also accounts for non-informative censoring, which occurs when either one or both conditions hold:

i) $\gamma_0^T = \ldots = \gamma_K^T = 0$;
ii) $\gamma_0^C = \ldots = \gamma_K^C = 0$;

In particular, in situation i) we obtain the Cox proportional hazards model without heterogeneity between patients in hazard for the time to event T, while in ii) the hazard function for T is still characterized by sources of heterogeneity which are not explained by the censoring mechanism.

Identifiability

General conditions for the identifiability of this modelling approach have been analysed in[21],[22] and[23].

Nevertheless, given the finite mixture nature of our approach, two additional conditions are needed to ensure its global identifiability[1] (see also[25] and[26]):

1. For every value of $(t, c, \mathbf{x^T}(t), \mathbf{x^C}(c))$, different values of H should result into different p.d.f.s:

$$\begin{aligned} & f\Big(t, c \mid \mathbf{x^T}(t), \mathbf{x^C}(c), H = j; \beta^T, \beta^C, \gamma_j^T, \gamma_j^C\Big) \\ & \neq f\big(t, c \mid \mathbf{x^T}(t), \mathbf{x^C}(c), H = k; \beta^T, \beta^C, \gamma_k^T, \gamma_k^C\big) \quad \text{for j} \neq \text{k} \end{aligned}$$

   and thus two different hazard functions for T or two different hazard function for C given their conditional independence;
2. $0 < \zeta_k < 1$ for $k = 0, \ldots, K$. This condition is needed because if for any k we have $\zeta_k = 0$, then γ_k^T and γ_k^C can take any value without affecting the mixture distribution.

However, the joint probability distribution of (T, C) outlined so far is not locally identifiable: by permuting the labels of H, we obtain exactly the same joint p.d.f.:

$$\begin{aligned} & \sum_{k=0}^{K} \zeta_k \left[\exp\left(-\int_0^t \mu^T\big(s \mid x^T(s), h; \beta^T, \gamma_k^T\big)\mathrm{d}s\right)\mu^T\big(t \mid x^T(t), h; \beta^T, \gamma_k^T\big)\right] \\ & \times \left[\exp\left(-\int_0^c \mu^C\big(s \mid x^C(s), h; \beta^C, \gamma_k^C\big)\mathrm{d}s\right)\mu^C\big(c \mid x^C(c), h; \beta^C, \gamma_k^C\big)\right] \\ & = \sum_{k=0}^{K} \zeta_{\rho(k)} \left[\exp\left(-\int_0^t \mu^T\Big(s \mid x^T(s), h; \beta^T, \gamma_{\rho(k)}^T\Big)\mathrm{d}s\right)\mu^T\Big(t \mid x^T(t), h; \beta^T, \gamma_{\rho(k)}^T\Big)\right] \\ & \times \left[\exp\left(-\int_0^c \mu^C\Big(s \mid x^C(s), h; \beta^C, \gamma_{\rho(k)}^C\Big)\mathrm{d}s\right)\mu^C\Big(c \mid x^C(c), h; \beta^C, \gamma_{\rho(k)}^C\Big)\right] \end{aligned}$$

where $\rho(k)$ represents any permutation of the indexing given by k, and $\sum_{k=0}^{K} \zeta_k = 1$.

This problem can be easily overcome by restricting the parameter space of the hazard function. In Section 3.2 we address this issue in the perspective of improving the efficiency of the Hamiltonian Monte Carlo sampler.

Inference

For the reasons described in Section 1, in this work we carry out our inferential exercise using Bayesian techniques. Let $p(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta \mid K)$ denote the K-dependent prior distribution of the parameter vector $(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta)$. The learning object is thus given by the following posterior distribution of the vector $(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta)$:

$$\begin{aligned} & p\big(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta \mid \text{data, K}\big) \propto p\big(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta \mid K\big)L\big(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta \mid \text{data, K}\big) \\ & = p\big(\gamma^T, \gamma^C, \zeta \mid K, \beta^T, \beta^C\big)p\big(\beta^T, \beta^C\big)L\big(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta \mid \text{data, K}\big) \end{aligned} \tag{5}$$

where $L(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta \mid \text{data, K})$ denotes the likelihood as a function of the parameters, given the data and K. In order to avoid the use of a cumbersome notation p will denote the density function of the parameters.

In equation (5) we remarked the dependence of the posterior distribution on K, since it determines the dimension of the parameter space, particularly for γ^T, γ^C and ζ.

Likelihood function

Let us first distinguish C from C^*, where the former denotes the time to informative censoring, and the latter denotes the time to non-informative censoring (e.g. administrative censoring). Let $T' = \min(\text{T, C, C}^*)$, $d_i = 1_{[t'_i = t_i]}$ and $d_i^* = 1_{[t'_i = c_i]}$. Thus, if a time to event is subject to non-informative censoring, then $d_i = d_i^* = 0$.

The latent factor H is never observed, hence the ith individual likelihood contribution must be marginalized with respect to H:

$$\begin{aligned} L_i\big(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta\big) &\propto f\big(t_i' \mid x_i^T\big([0, t_i']\big), x_i^C\big([0, t_i']\big); \beta^T, \gamma^T, \beta^C, \gamma^C, \zeta\big) \\ &= \sum_{k=0}^{K} \zeta_k \left[\exp\left(- \int_0^{t_i'} \mu_0^T(s) \exp\big(\beta^{T'} x_i^T(s) + \gamma_k^T\big) \mathrm{d}s \right) \big(\mu_0^T\big(t_i'\big) \exp\big(\beta^T x_i^T(t_i') + \gamma_k^T\big)\big)^{d_i} \right] \\ &\times \left[\exp\left(- \int_0^{t_i'} \mu_0^C(s) \exp\big(\beta^{C'} x_i^C(s) + \gamma_k^C\big) \mathrm{d}s \right) \big(\mu_0^C\big(t_i'\big) \exp\big(\beta^{C'} x_i^C(t_i') + \gamma_k^C\big)\big)^{d_i^*} \right] \end{aligned} \tag{6}$$

where $\gamma_0^\ell = 0$ and $x_i^\ell(s)$ for $\ell \in \{T, C\}$ is the set of covariates value at time $s \in [0, t_i']$.

In equation (6) the covariates are assumed as continuously observed throughout the whole time span $[0, t_i']$. In practice, covariates are observed at discrete time points, and several assumptions can be made for their value at intermediate time points. For example, in the simulation study (Section 4) and in the empirical analysis (Section 5) the covariates are assumed constant between two observation times, and equal to the previously observed value.

The resulting likelihood function of the parameters, given the data $L(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta \mid \text{data}) = \prod_{\mathrm{i}} \mathrm{L}_{\mathrm{i}}(\beta^{\mathrm{T}}, \gamma^{\mathrm{T}}, \beta^{\mathrm{C}}, \gamma^{\mathrm{C}}, \zeta)$ is the sum of $(K+1)^n$ terms, each corresponding to a realization of the latent factor for each unit in the sample.

Prior specification and model reparametrization

In general, the Bayesian analysis of a mixture model is particularly challenging. For this reason, the prior should be chosen accurately, and the model needs to be parametrized in a way that makes the sampling process efficient.

First of all, we assume for convenience that the vectors $\beta^T, \beta^C, \gamma^T, \gamma^C$ and ζ are pairwise independently distributed, that is:

$$p\big(\beta^T, \gamma^T, \beta^C, \gamma^C, \zeta \mid K\big) = p\big(\beta^T\big)p\big(\beta^C\big)p\big(\gamma^T \mid K\big)p\big(\gamma^C \mid K\big)p\big(\zeta \mid K\big) \tag{7}$$

where each density depends on some hyperparameters.

The specification of the prior distribution for the regression coefficient parameters β^T and β^C does not represent a problem, hence it is possible to specify a non-informative prior in order to reduce the extent of subjective assumptions on the inference.

The prior distribution of γ^T, γ^C and ζ must be specified with more care. The use of prior distributions which are at least weakly informative is appropriate, in order to enhance the efficiency of the sampler.

For example, with the use of an exchangeable prior on $(\gamma_1^T, \ldots, \gamma_K^T)$, $(\gamma_1^C, \ldots, \gamma_K^C)$ and of a uniform prior on ζ^2, then their posterior distribution will maintain the permutation invariance induced by the mixture likelihood ([27] and[28]).

This multimodality hinders the efficiency of the sampler, because it would take a large number of iterations to fully explore the posterior distribution. Indeed, the sampler is likely to get stuck in an area of the parameter space corresponding to one mode of the posterior distribution (see[29]).

In Bayesian inference the identifiability of a mixture distribution turns out to be more crucial. In Section 2.2 we mentioned how a constraint allows for a model to be identifiable. In our case we also need to follow a strategy which allows for an efficient sampling for the posterior distribution.

For the joint distribution of equation (4), we need to place an ordering constraint on the set $[(\gamma_1^T, \gamma_1^C), \ldots, (\gamma_K^T, \gamma_K^C)]$, which can be done by ordering either $(\gamma_1^T, \ldots, \gamma_K^T)$ or $(\gamma_1^C, \ldots, \gamma_K^C)$. For example, we can set:

$$0 < \gamma_1^T < \gamma_2^T < \ldots < \gamma_K^T \tag{8}$$

where we take into account also the component corresponding to $\gamma_0^T = 0$. An account of the geometric interpretation of this constraint can be found in[28]. This constraint also turns exchangeable priors into nonexchangeable ones due to the restriction of the parameter space into its subset, singling out one mode of the posterior distribution, as induced by the label permutation. Conversely, the exchangeability in the prior distribution will be inherited also by the posterior distribution.

However, when the mixture components are not well separated (i.e. these are very close to one another in value), then also the posterior distribution will be affected by the imposed constraint. For example, suppose we set an ordering as in equation (8), and there is a nonnegative probability that $\gamma_1^T > \gamma_2^T$ (all else being equal, which is more likely if γ_1^T is very close to γ_2^T), then the true posterior distribution cannot be captured by the posterior distribution satisfying such constraint.

In the extreme case where the mixture components overlap, then the posterior distribution will be more difficult to explore ([28]), as may occur when the value of K is larger than actually needed to explain the heterogeneity among the units. In this case the components in excess will collapse into those which are necessary and have been already included in the model.

This justifies our inferential strategy of analyzing increasingly complex models in terms of needed mixture components, as we describe in Section 3.3.

Choice of the number of mixture components

The value of K is inferred by solving a model selection problem, where models with different values of K are separately fitted and then compared to one another using appropriate information criteria. Algebraic geometry tools have been proven useful in order to understand the asymptotic behavior of the posterior distribution of the parameters when dealing with singular statistical models. For example,[11] proposes the Widely Applicable Information Criterion (WAIC), which generalizes the AIC to the analysis of singular models.

The WAIC is computed as follows (notation simplified for ease of exposition):

$$\mathrm{WAIC} = -\sum_{i=1}^{n} \log\left[\frac{1}{M}\sum_{m=1}^{M} f\left(t_i' \mid x_i^T(s), x_i^C(s), d_i, d_i^*; \beta^{T(m)}, \beta^{C(m)}, \gamma^{T(m)}, \gamma^{C(m)}, \zeta^{(m)}\right)\right] + p_{\mathrm{WAIC}};$$

$$p_{\mathrm{WAIC}} = 2\sum_{i=1}^{n}\left[\log\left(\frac{1}{M}\sum_{m=1}^{M} f\left(t_i' \mid x_i^T(s), x_i^C(s), d_i, d_i^*; \beta^{T(m)}, \beta^{C(m)}, \gamma^{T(m)}, \gamma^{C(m)}, \zeta^{(m)}\right)\right)\right.$$
$$\left. - \frac{1}{M}\sum_{m=1}^{M} \log f\left(t_i' \mid x_i^T(s), x_i^C(s), d_i, d_i^*; \beta^{T(m)}, \beta^{C(m)}, \gamma^{T(m)}, \gamma^{C(m)}, \zeta^{(m)}\right)\right]$$

where M is the number of sampled values from the posterior distribution of the parameters, and p_{WAIC} is a penalization term which can be interpreted as the effective number of parameters, and measures the fluctuation of the posterior distribution (see[16]).

Our strategy is to first analyze the model with $K = 0$ (T and C independently distributed given $x^T(\cdot)$ and $x^C(\cdot)$) and then we fit models with increasing K until the WAIC does not decrease, as we aim at its minimization.

Similar information criterion have been proposed to generalize the BIC, such as the singular BIC of[12] and the Widely Applicable Bayesian Information Criterion (WBIC) of[30]. The former is of harder implementation in practice, while we found in a simulation study (not discussed here) that the WBIC shows a slightly better performance than the WAIC in selecting the true model, particularly when the mixture components are well separated. The WAIC is hereby preferred over the BIC generalization for two practical reasons: i) it provides a measure of model dimension through p_{WAIC}; ii) it turns out to be more practical since it allows to use the same output from the posterior sampling process.

Estimation

For the estimation of the posterior distribution of equation (5) we use the Hamiltonian Monte Carlo (HMC) sampler (see[31]). This algorithm belongs to the general Metropolis-Hastings family. HMC uses Hamiltonian dynamics which allow for a faster exploration of the parameter space. It turns out to be more efficient than more traditional Gibbs and random-walk Metropolis samplers for posterior distributions showing high correlations among the parameters, as it occurs for mixture distributions.

The HMC sampler is implemented by using the Stan software package ([32]) and its R interface ([33]), by means of the package `rstan`. In its default implementation Stan uses the No-U-Turn sampler of[34] which allows for an automatic tuning of the sampler. Furthermore, the `rstan` can automatically initialize the sampling process.

Simulation study

Set up

Simulation studies are carried out to assess the performance of the methodology outlined in this work. We focus on the posterior distribution of the parameters β^T, β^C, γ^T and γ^C, and thus on whether the WAIC allows to chose the true model. We look at whether the 95% credible intervals of the posterior distribution include the true value of the parameters, and we analyse whether the posterior mean can be considered as an unbiased point estimate of the parameters.

We compare these results with the parameter estimates and the 95% coverage intervals obtainable from the Cox proportional hazards model. This gives us the opportunity to quantify the need for our method in certain settings.

For each simulation run, we generate samples of 500 potentially censored lifetimes.

For each patient we assume the vector of time-varying binary covariates $X(t) = X^T(t) = X^C(t) = (X_1(t), X_2(t))$ (for example treatment and disease) with the probability distribution shown in Table 1.

The binary covariate variables $(X_1(t), X_2(t))$ are independent throughout time.

We consider an observational period of 8 years, and a follow up period of 4 years for simplicity such that the researcher observes $X(t)$ for $t = 0, 4, 8$. In medical practice, these values can change at very high frequency over time, for example whether a patient has a high or low blood pressure, or the time-dependent treatment, although applied statistical analyses contain simplifications to overcome the issue. In this simulation study we assume that $X(t)$ is constant between one observation and the other.

For each unit, H is generated from a Mult$(1, \zeta)$ distribution.

For $K = 1$ we consider two scenarios: the case of well separated mixture components (WS), and the case of poorly separated mixture components (PS).

In the WS scenario T and C are generated under the following hazard functions:

$$\mu^T\big(u \mid x^T(u), h; \beta^T, \gamma^T\big) = \exp\big(-1.2 - 0.9x_1(u) + 0.8x_2(u) + 2h_1\big) \tag{9}$$

$$\mu^C\big(u \mid x^C(u), h; \beta^C, \gamma^C\big) = \exp\big(-2 - x_1(u) + 0.5x_2(u) + 3.5h_1\big) \tag{10}$$

Conversely, in case of poorly separated mixture components, we specify:

$$\mu^T\big(u \mid x^T(u), h; \beta^T, \gamma^T\big) = \exp\big(-0.2 - 0.9x_1(u) + 0.8x_2(u) + 0.5h_1\big) \tag{11}$$

$$\mu^C\big(u \mid x^C(u), h; \beta^C, \gamma^C\big) = \exp\big(-0.5 - x_1(u) + 0.5x_2(u) + h_1\big) \tag{12}$$

For the case of two mixture components we set $\zeta = (0.65, 0.35)$ and γ^C is chosen in order to keep a stochastic ordering in the hazard function of the censoring time due to higher values of H.

When $K = 2$, then T and C are generated according to the following hazard functions:

$$\mu^T\big(u \mid x^T(u), h; \beta^T, \gamma^T\big) = \exp\big(-1 + 0.3x_1(u) - x_2(u) + 3h_1 + 2h_2\big) \tag{13}$$

$$\mu^C\big(u \mid x^C(u), h; \beta^C, \gamma^C\big) = \exp\big(-1.6 - 0.7x_1(u) + 0.5x_2(u) + 2.5h_1 + 4h_2\big) \tag{14}$$

and $\zeta = (0.48, 0.32, 0.2)$. In this case, we consider only the case of well separated mixture components due to the smaller sample size.

The analysis of models with more than three mixture components requires ideally larger samples than those analysed in this work, while the case $K = 0$ can be more efficiently fitted with a Cox proportional hazard model. For each value of K we generate 100 datasets. In all cases we ensure that the datasets have around 40% censored units and a negligible percentage of type I censoring cases.

Drawing a value for (T,U)

A general property of survival models is that the integrated hazard function $\int_0^t \mu(s)\mathrm{d}s$ is a random variable with Exponential(1) distribution.

Table 1. Joint probability distribution of $(X_1(t), X_2(t))$.

		$X_2(t)$		
		0	1	Mar. $X_1(t)$
$X_1(t)$	0	0.37	0.23	0.6
	1	0.33	0.07	0.4
	Mar. $X_2(t)$	0.7	0.3	

Therefore, conditional on a standard uniform random variable $Z \sim \text{Unif}(0, 1)$, a set of covariates $X(t)$ and a factor H, a sampled value of T, denoted by t, is given by the solution of the following equation:

$$-\log z = \int_0^t \mu^T\big(s \mid x(s), h; \beta^T, \gamma^T\big)\mathrm{d}s \tag{15}$$

Given the set-up described in Section 4.1 and the hazard function specification of equations (9)-(14), then equation (15) has a closed form solution. In the same way we can sample a value for the censoring time C.

Then, once t and c are obtained, a value for the triplet (t', d, d^*) is calculated as follows:

$$t' = \min(\text{t, c}) \tag{16}$$

$$d = 1_{[t \leq c \leq 8]} \tag{17}$$

$$d^* = 1_{[c < t \leq 8]} \tag{18}$$

Implementation

We use weakly informative prior distributions for each parameter, instead of fully non-informative ones in order to foster the convergence of the sampler. In this way, the posterior distribution will not be strongly affected by prior assumptions. Indeed, with enough observations, the effect of the prior should be negligible ([16]).

Furthermore, we assume that parameters are *a priori* independently distributed.

We model the baseline hazard function of T and C using a piecewise constant hazard function as follows:

$$\mu_0^\ell(t) = \begin{cases} \exp\big(\lambda_{00}^\ell\big) & \text{ift} < 4 \\ \exp\big(\lambda_{01}^\ell\big) & \text{ift} \geq 4 \end{cases}$$

where $\ell \in \{T, C\}$, in order to ease the representation, considering a time span divided into two sub-periods of 4 years.

The posterior modeling involves $\lambda_0^\ell = (\lambda_{00}^\ell, \lambda_{01}^\ell)$, since in this way we do not need to constrain the parameter space, and the sampler works more efficiently.

The piecewise constant baseline hazard function represents a useful yet flexible specification, mainly if the number of knots (or intervals) increases. Alternative models can be given for example by a Gamma Process for the integrated baseline hazard function (see[35]), or the use of splines ([36]). Further alternatives are discussed in the book of[37].

For the prior distribution we assume that $\lambda_0^T, \lambda_0^C, \beta^T$ and β^C are vectors of pairwise independent and normally distributed variables with mean 0 and variance equal to 10. When fitting a model with $K = 1$ we assume that $\gamma_1^T \sim N(0.5, 100)$, $\gamma_1^C \sim N(1, 100)$ truncated at the lower bound of 0 and $\zeta \sim$ Dirichlet(30, 20). When $K = 2$, the prior of γ_1^T and γ_1^C are as before, while $\gamma_2^T \sim N(1, 100)$, $\gamma_2^C \sim N(2, 100)$ truncated at the lower bound of γ_1^C and $\zeta \sim$ Dirichlet(21, 17, 12). In a similar fashion, when fitting a model with $K = 3$ we assume $\gamma_3^T \sim N(2, 100)$, $\gamma_3^C \sim N(3, 100)$ truncated at the lower bound of γ_2^C and $\zeta \sim$ Dirichlet(20, 15, 10, 5). In all cases, when choosing the Dirichlet distribution parameters, we kept a prior sample size (the sum of the Dirichlet distribution parameters) of 50.

The HMC sampler is run for 5,000 iterations and the first half draws are discarded in order to allow for burn-in and fine-tune the No-U-Turn sampler. Indeed, given our simulation settings, 5,000 iterations are sufficient to allow for the chains to mix and obtain a stationary posterior distribution, which ensures convergence of the sampler.

Results

The results in Table 2 for $K = 1$ show that the 95% credible intervals from the posterior distribution of the parameters include their true value with a coverage close to normal when fitting the true model $K = 1$ and the mixture components are well separated. In case of lesser separated mixture components, the 95% credible intervals still show a coverage larger than 90% for β^T, β^C and ζ. However, the coverage sensibly decreases for γ^T and γ^C, whose posterior mean turns out to be divergent from their true values.

When the mixture components are well separated, the 95% credible intervals have a coverage below nominal for almost all parameters when fitting the model assuming that T and C are independently distributed (K=0): β_1^T and β_2^T have a coverage of 83% and 95% respectively, and β_1^C and β_2^C are included in the credible intervals with 64% and 82% probability respectively. As expected, in case of poorly separated mixture components, without a very large sample size, the model

Table 2. Parameter estimates with the Cox proportional hazard model (Cox PH), posterior mean estimate for the models with $K = 0, 1, 2$, coverage of the 95% confidence interval (for the Cox model) and coverage of the 95% credible intervals.

Par.	True v.	Par. est	Post. mean			95% CI Cov.	**95**% Cred. int. Cov.		
		Cox PH	K=0	K=1	K=2	Cox PH	K=0	K=1	K=2
Well separated mixture components									
β_1^T	−0.9	−0.74	−0.79	−0.89	−0.90	75	83	94	95
β_2^T	0.8	0.68	0.78	0.80	0.83	85	95	98	94
γ_1^T	2	–	–	1.97		–	–	95	–
β_1^C	−1	−0.55	−0.75	−1.02	−1.04	22	64	92	93
β_2^C	0.5	0.21	0.43	0.49	0.47	56	82	93	93
γ_1^C	3.5	–	–	3.50	–	–	–	98	–
ζ_0	0.65	–	–	–	0.65	–	–	97	–
ζ_1	0.35	–	–	–	0.35	–	–	97	–
Poorly separated mixture components									
β_1^T	−0.9	−0.90	−0.88	−0.89	−0.89	92	93	92	94
β_2^T	0.8	0.76	0.83	0.82	0.85	89	94	95	94
γ_1^T	0.5	–	–	−4.12		–	–	67	–
β_1^C	−1	−0.89	−0.96	−0.97	−0.98	92	96	94	96
β_2^C	0.5	0.45	0.49	0.50	0.46	94	90	90	91
γ_1^C	1	–	–	2.00	–	–	–	79	–
ζ_0	0.65	–	–	–	0.61	–	–	99	–
ζ_1	0.35	–	–	–	0.39	–	–	99	–

with $K = 0$ tends to show a coverage close to nominal for the parameters β^T and β^C. Indeed, the true model tends to be closer to the model with $K = 0$.

The results for the posterior mean follow from those of the credible intervals. We see that the regression parameters β^T and β^C are in line with respect to their true value when $K = 1, 2$ and the mixture components are well separated. Same applies for the baseline hazard (not shown) when fitting the true model with $K = 1$, as well as for γ^T, γ^C and ζ.

Analogous results are obtained for the model with $K = 2$ as shown in Table A1 in Appendix A.1. For this particular case, with a sample of 500 units we note that the posterior mean of γ_2^T is different compared to its true value, despite the credible intervals show a coverage of 92%. A closer inspection of this result showed that this is the consequence of the small number of events for those units with $H = 2$ (around 5% of the observations), which causes γ_2^T to be estimated with larger uncertainty.

Table 3 shows that when the mixture components are well separated in all cases the WAIC excludes the lack of heterogeneity, ruling out the model with $K = 0$. In 90 cases out of 100 it selects the model with K=1, while in 10 cases it selects the model with $K = 2$. In this latter case, the WAIC for the model $K = 2$ is never higher than 2 compared to the model with $K = 1$. This means that if we chose the more parsimonious model with $K = 1$, we do not lose much information. Clearly, these results depend on the values of γ^T and γ^C, which we purposely chose in order to ensure that the mixture components are well separated. Conversely, when the mixture components are poorly separated, in case of smaller sample sizes the models with $K = 0, 1, 2$ tend to behave similarly. Indeed, when the WAIC picks the model with $K = 0$, then it is

Table 3. Number of times out of 100 the WAIC selects the true model based on *K*.

True	Sample	Mixture	Fitted models			
model (K)	size	sep.	K=0	K=1	K=2	K=3
$K = 1$	500	PS	41	**19**	40	0
$K = 1$	500	WS	0	**90**	10	0
$K = 2$	500	WS	0	54	**39**	7
$K = 2$	1,000	WS	0	18	**60**	22
$K = 2$	5,000	WS	0	7	**83**	10

never larger than 1, compared to the true model with $K = 1$. Similarly, when the WAIC picks the model with $K = 2$, the difference with the WAIC for the true model is never larger than 3.2.

Given our choice of the parameters in equations (13)–(14) related to the case $K = 2$ we could see that the WAIC chooses the true model in 39 cases, while in 54 cases it chose the model with $K = 1$ and in 7 cases the model with $K = 3$. The model with $K = 0$ is always ruled out. When the model with $K = 3$ is chosen the difference in WAIC is never higher than 1.79, and thus we reach the same conclusions as the previous cases.

These results do not necessarily mean that our approach to model selection has a poor performance as the true K increases: the more complex the model, the larger the sample size should be in order to capture the heterogeneity in the units. Indeed when using a sample size of 1,000 and of 5,000 units, then the true model is selected in 60 and 83 cases out of 100 respectively. In addition, as aforementioned, also the true values of the parameter γ^T and γ^C play a fundamental role to capture heterogeneity among the units.

This simulation study showed how the model is always capable to spot the heterogeneity in the distribution of T induced by informative censoring. Another key point is that failing to account for this heterogeneity yields a bias in the estimation of the regression coefficients, as shown by their smaller coverage. The probability of selecting the true model increases with sample size, as we can expect in highly parametrized models as analysed in this work. Parsimony is also preserved, since even when fitting a larger model which shows a better performance in terms of WAIC, then selecting a smaller model does not result in a great loss of information. Compared to the use of the Cox proportional hazard model, the results of this work are even more robust if we consider that in our model we needed to specify a baseline hazard function.

Analysis of ACTG 175 dataset

The AIDS Clinical Trial Group (ACTG) 175 study ([17]) is a double-blind randomised clinical trial where a total of 2,467 adults infected with HIV type I and CD4 cell counts between 200 and 500/mm^3 are randomly assigned to four different treatments: i) zidovudine, ii) didandosine, iii) zidovudine plus didandosine and iv) zidovudine plus zalcitabine.

Enrolment run from December 1991 until October 1992, while patients were scheduled to stay under analysis until November 1994. In particular, they are examined at weeks 2, 4, 8 and every 12 weeks afterwards.

For each patient we could observe baseline covariates, such as age, gender, ethnicity, CD4 count, Karnofsky score, prior use of antiretroviral therapy, haemophilia and whether the HIV was symptomatic or not; as well as some other information like the primary endpoint, and the reason for drop out from the study[3] . CD4 cell counts are checked at the initial visit and then from week 8 onwards.

The primary endpoints (the time to the event of interest we want to model in this work) of the study were a 50% decrease in CD4 cell counts with respect to the baseline, AIDS or death. At the end date of the observational study there are some patients whose primary end point did not occur, hence these can be considered as non-informatively type 1 censored.

However, some other patients can drop out from the study earlier than the end date without reaching the primary endpoint for some reasons, such as toxicity of the therapy, request of the patients themselves or of the investigator, while some others just do not show up at the next planned visit (loss to follow up). In these cases it is reasonable to assume that these drop out causes are related to the primary endpoint, since for example a patient may ask to discontinue a therapy because she feels better, therefore the death event is less likely to occur (other things being equal). In this case, the Kaplan-Meier estimate of the survival function (which considers censoring as non informative) is likely to be pessimistic about patient survival.

Similar to the work of[6] and[39] we focus on the analysis of those patients treated with zidovudine only. We thus have 614 patients of which 195 experienced the primary endpoint (151 had 50% CD4 reduction, 14 died and 30 developed AIDS).

Hence, of the remaining 419 patients, 183 are Type I censored and 236 are considered as subject to informative censoring.

Furthermore,[17] observed that throughout the study ”younger patients, those reporting injection-drug use and those with lower CD4 cell counts, lower Karnofsky score and symptoms of HIV infection at enrollment are more likely to discontinue the treatment before the study ends”.

For illustration we consider the following covariates vector for T and C:

$$X^T(t) = \text{CD4(t)}$$

$$X^C(t) = \big(\text{age, iv, kar, sym, CD4(t)}\big)$$

We chose $X^T(t)$ based on the WAIC, while $X^C(t)$ is chosen on the basis of the aforementioned considerations of[17]. $X^C(t)$ is the same vector of covariates chosen in[6] and[39].

In particular, *age* denotes the age of the patient at time of randomization, *iv* is an indicator of intravenous drug use, *kar* indicates the Karnofsky score, *sym* is a binary variable indicating the presence of symptoms of HIV infection and CD4(t) indicates the CD4 cells count at each visit (t). We assume for simplicity that the CD4 cell count is constant between one visit and the other.

From a look at the Kaplan-Meier curve of the integrated hazard function, shown in Figure 1 we could have a rough idea about the shape of the baseline hazard function which seems approximately piecewise linear for both T and C. For this reason we specify a piecewise constant baseline hazard function for $\mu_T(\cdot)$ and $\mu_C(\cdot)$ with two breakpoints. As stated in Section 4.3 other specifications for the non-parametric baseline hazard function can be possible. For this dataset we noted that the use of a Gamma process would rather slower the HMC sampler without returning better results.

The breakpoints have been chosen by fitting a piecewise linear function with two breakpoints where the cumulative hazard function obtained from the Kaplan-Meier estimator is regressed against the time to event using the package `segmented` ([40]). The breakpoint vector for μ_0^T is (368.5, 1006.2), while the vector for μ_0^C is (841, 958.2).

In a similar fashion with respect to the simulation study we chose weakly informative prior distributions. Hence, we assume that λ_0^T, λ_0^U, β^T and β^C are vectors of independent and normal random variables with mean 0 and standatd deviation equal to 10. γ^T and γ^C are chosen in the same way as in the simulation study, while assuming a prior sample size for ζ equal to 25 for all fitted models. For $K = 1, 2, 3$ different prior distribution have been specified for ζ. This does not exclude the possibility for the researcher to use expert judgement when specifying the prior distribution. The choice of the best model is again based on the WAIC.

For each value of K we run 4 parallel chains, each for 20,000 iterations (10,000 used as warm-up).

In addition, we compare these results with those obtainable from fitting a Cox proportional hazard model, which assumes independence between T and C.

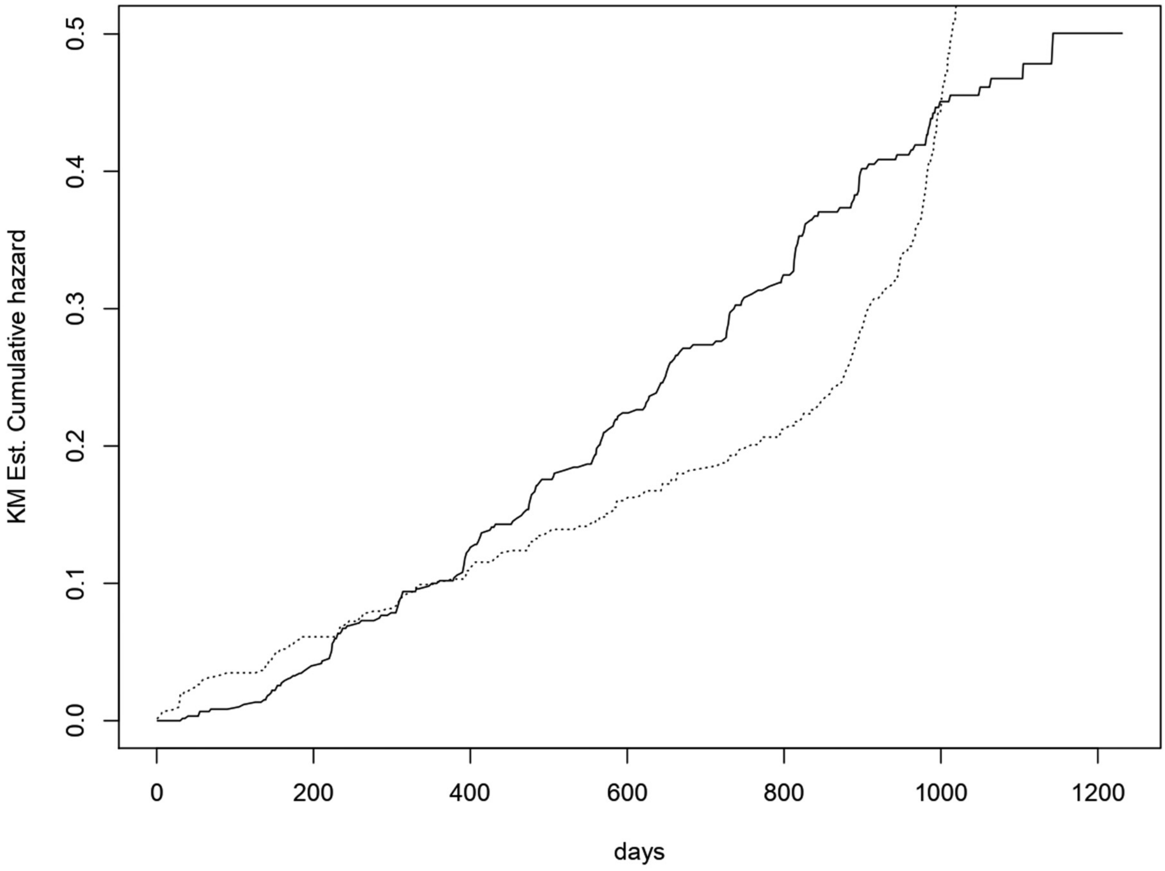

Figure 1. Kaplan Meier estimate of the cumulative hazard function for T (solid line) and for the censoring (dotted line).

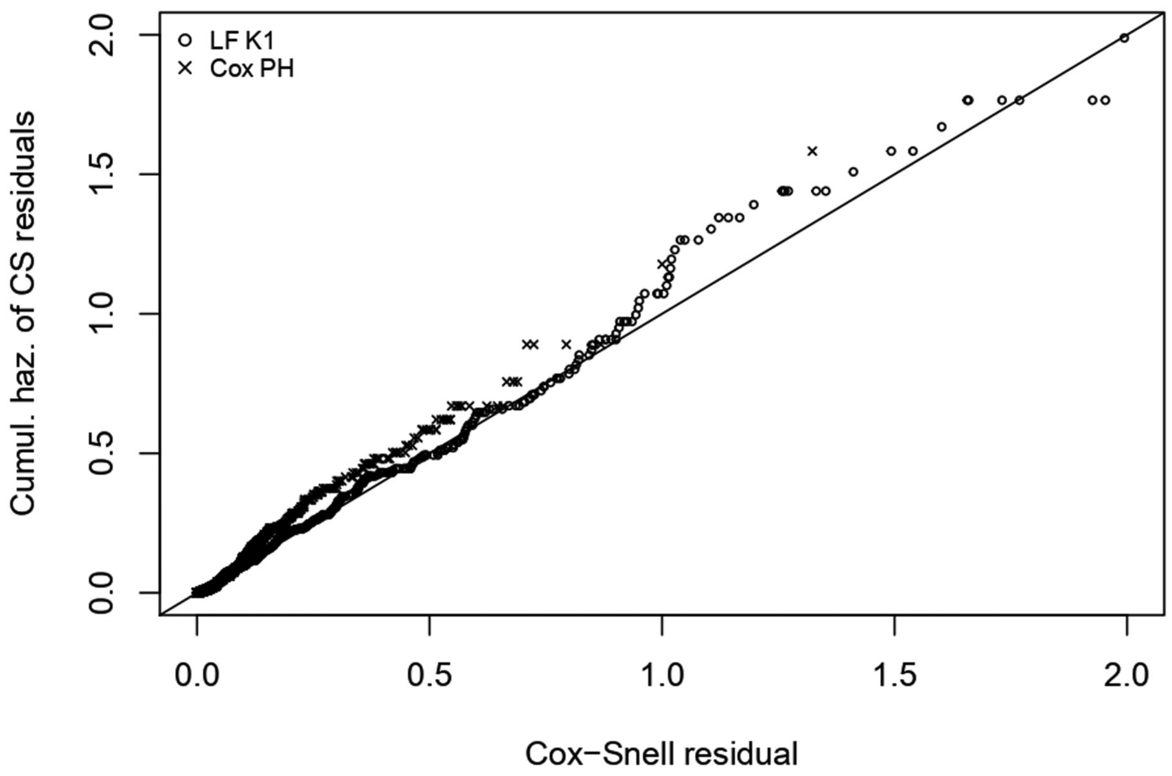

Figure 2. Cox-Snell residuals for the latent factor model with $K = 1$ (LF K1, o) and for the Cox PH model (x) plotted against their cumulative (or integrated) hazard function.

Results

We observe that for the analysed models the HMC sampler converged towards the posterior distribution, as we can see for example from the traceplots of Figure A1, where we showed only the first two chains in Section A.2.1, and the $\widehat{R}$ statistic of[41] for all parameters which is always equal to 1. The marginal posterior density of each parameter is nearly symmetric, except for γ_1^C due to the chosen local identifiability constraint (Figure A2).

The WAIC leads to the choice of the model with $K = 1$ (Table 5), whose results are shown in Table 4. As emphasized in the simulation study of Section 4 the true value of K may be higher than 1 due to the relatively small sample size, although we did not have a great loss of information. As we can see, the WAIC progressively increase with K after hitting its lowest point at the value of 1.

For all values of K we see that lower time to primary endpoints are associated with lower CD4 cell counts, coherently to our expectations. This is because it is an indicator of progression of the HIV and of immunologic health.

Again, when analysing the time to censoring, we can see that our results are consistent with the observations of[17] mentioned in the previous section, hence with the evidences obtained by[6].

The standard errors of the parameter estimates obtained with the Cox PH model are smaller than the standard deviations of the posterior distribution of the parameters from our Bayesian modelling approach. This is presumably due to the fact that the Cox model tends to yield lower uncertainty about parameter estimates than the joint model developed in this work since the latter has more parameters, as we specify a baseline hazard and include a latent factor.

The covariates intravenous drug use and the presence of symptoms do not show statistical significance, which may be explained by the role of CD4 cell counts. For example,[42] in the analysis of the progression of the HIV between hard drug users and other subjects in the Women's Interagency HIV Study, noted that hard drug users tend to drop out from the study earlier, and that those subjects have lower mean of CD4 cell counts. Thus the effect of intravenous drug use may be mediated by CD4 cell count. Additionally, the CD4 count may be a confounder for the effect of symptoms and the direct effect of symptoms is relatively small when corrected for CD4 cell count.

For this particular dataset the summaries of the posterior distribution of β^T and β^C from the model with $K = 1$ are very close to those for the models with $K = 0, 2, 3$, as we can see in Tables A2–A4 in Appendix A.2.2, although we can see some major differences with reference to the baseline hazard functions (from 16 to 40% in absolute value), $\beta^T_{CD4(t)}$

(25% higher than the same parameter for the model with $K = 0$) and $\beta^C_{CD4(t)}$ (34% lower). As we could see also from the simulation study, the assumption of independence may lead to considerable risk of wrong parameter estimates following model misspecification. When comparing the results of the latent factor model with $K = 1$ with the Cox proportional hazard model which leaves the baseline hazard function unspecified (Table 4), once again we see that the heterogeneity may have a considerable effect on the value of parameter estimates which may be biased if this further heterogeneity is not taken into account. For example, the Cox model underestimates the effect of the CD4 cell counts.

The parameter γ^T_1, has a marginal posterior distribution such that values around zero are very unlikely. This means that there is further heterogeneity in T which may be captured from other factors. When coupled with the distribution of γ^C_1 (Figure A2) we can see that this heterogeneity may be due to the presence of informative censoring (posterior mean equal to 1.07). A further sign of convergence is that large probability mass lies far from the boundary of the sample space, as determined by the constraint.

In addition, we compare these two models in terms of goodness of fit, which can help in the choice between these two models. At this purpose we look at the Cox-Snell residuals, r^{CS}_i ([43]), defined as the integrated hazard function. A well known result in mathematical statistics states that these residuals have an Exp(1) distribution if the model has been properly specified.

For the Cox PH model, these can be derived as a by product of the estimation process by using the package `survival` ([44]), while for the latent factor model of this work, r^{CS}_i is calculated as follows:

$$\begin{aligned} r^{CS}_i &= \mathbb{E}\big[-\ln S_T\big(t'_i \mid x^T_i(\cdot), x^C_i(\cdot), c_i, d^*_i, \lambda^T, \lambda^C, \beta^T, \beta^C, \zeta\big)\big] \\ &= \mathbb{E}\left[-\ln\sum_{k=0}^{K} A_k + \ln\sum_{k=0}^{K} B_k\right] \end{aligned} \tag{19}$$

where

$$A_k = \exp\left[-\int_0^{t'_i} \mu^T\big(s \mid x^T_i(s), h; \beta^T, \gamma^T\big)\mathrm{d}s\right] B_k \tag{20}$$

$$B_k = \exp\left[-\int_0^{t'_i} \mu^C\big(s \mid x^C_i(s), h; \beta^C, \gamma^C\big)\mathrm{d}s\right] \mu^C\big(t'_i \mid x^C_i\big(t'_i\big), h; \beta^C, \gamma^C\big)^{d^*_i} \tag{21}$$

In other words, unlike the Cox PH model where we have a point estimate for the integrated hazard function, we average the integrated hazard function for the latent factor model over the draws from the posterior distribution. In addition, since we jointly model the time to event and the censoring time, we need to consider the probability distribution of T conditional on

Table 4. Cox proportional hazard model estimate of the parameters (Cox PH) and summary of the posterior distribution of the parameters for the model with $K = 1$.

Par.	Cox PH	Post. mean	Post. quantile		
	(est. s.err.)	(st. dev.)	2.5-th%	50-th%	97.5-th%
λ^T_{00}	–	−3.85(0.32)	−4.48	−3.85	−3.21
λ^T_{01}	–	−3.17(0.35)	−3.84	−3.18	−2.45
λ^T_{02}	–	−3.78(0.59)	−4.96	−3.77	−2.66
λ^C_{00}	–	−4.40(1.03)	−6.53	−4.39	−2.45
λ^C_{01}	–	−3.08(1.04)	−5.20	−3.06	−1.13
λ^C_{02}	–	−1.93(1.03)	−4.03	−1.91	0.03
$\beta^T_{\mathrm{CD4(t)}}$	−0.0104(0.0007)	−0.0139(0.00104)	−0.016	−0.0139	−0.0119
γ^T_1	–	−2.03(0.33)	−2.68	−2.03	−1.36
β^C_{age}	−0.046(0.0089)	−0.05(0.01)	−0.07	−0.05	−0.04
β^C_{iv}	0.74(0.1678)	0.76(0.19)	0.39	0.76	1.12
β^C_{kar}	−0.019(0.0107)	−0.03(0.01)	−0.05	−0.03	−0.01
β^C_{sym}	0.25(0.1770)	0.18(0.19)	−0.19	0.19	0.55
$\beta^C_{\mathrm{CD4(t)}}$	−0.00069(0.0005)	−0.00085(0.0005)	−0.0019	−0.00081	−0.000057
γ^C_1	–	1.07(0.62)	0.08	1.02	2.43
ζ_0	–	0.45(0.08)	0.29	0.45	0.61
ζ_1	–	0.55(0.08)	0.39	0.55	0.71

Table 5. WAIC and effective number of parameters (p_{WAIC}) for the models with $K = 0, 1, 2, 3$.

	K=0	K=1	K=2	K=3
WAIC	3,462.76	**3, 456.4**	3,457.34	3,458.13
p_{WAIC}	12.30	13.83	14.90	15.48

the censoring time (other than the covariates).

In order to check whether r_i^{CS} has Exp(1) distribution, we then calculate the Kaplan-Meier estimate of such residuals, and we plot r_i^{CS} against their Kaplan-Meier estimate of the integrated hazard function, as shown in Figure 2.

We see that the Cox-Snell residuals of the latent factor approach of this work follows more closely the 45-degrees straight line, compared to the Cox PH model, meaning that the former returns a better fit for the distribution of the time to event T.

A by-product of this analysis is the possibility to analyse the profile of the two subgroups (since K=1) arising from this modelling approach. Using Bayes' theorem we can calculate the posterior distribution of H for each patient q_{ih}:

$$\begin{aligned} q_{ih} = \mathrm{Pr}\big(\mathrm{H = h} \mid \mathrm{t}_\mathrm{i}', \mathrm{x}_\mathrm{i}^\mathrm{T}(\cdot), \mathrm{x}_\mathrm{i}^\mathrm{C}(\cdot)\big) &= \frac{\mathrm{Pr(H = h)Pr}\big(\mathrm{t}_\mathrm{i}', \mathrm{x}_\mathrm{i}^\mathrm{T}(\cdot), \mathrm{x}_\mathrm{i}^\mathrm{C}(\cdot) \mid \mathrm{H = h}\big)}{\sum_{k=0}^{K} \mathrm{Pr(H = k)Pr}\big(\mathrm{t}_\mathrm{i}', \mathrm{x}_\mathrm{i}^\mathrm{T}(\cdot), \mathrm{x}_\mathrm{i}^\mathrm{C}(\cdot) \mid \mathrm{H = k}\big)} \\ &= \frac{\zeta_h A_h'}{\sum_{k=0}^{K} \zeta_k A_k'} \end{aligned} \tag{22}$$

where

$$\begin{aligned} A_h' =& A_h \mu^T\big(t_i' \mid x_i^T\big(t_i'\big), h; \beta^T, \gamma^T\big)^{d_i} \\ =& \exp\left[-\int_0^{t_i'} \mu^T\big(s \mid x_i^T(s), h; \beta^T, \gamma^T\big)\mathrm{d}s\right] \mu^T\big(t_i' \mid x_i^T\big(t_i'\big), h; \beta^T, \gamma^T\big)^{d_i} \\ &\times \exp\left[-\int_0^{t_i'} \mu^C\big(s \mid x_i^C(s), h; \beta^C, \gamma^C\big)\mathrm{d}s\right] \mu^C\big(t_i' \mid x_i^C\big(t_i'\big), h; \beta^C, \gamma^C\big)^{d_i^*} \end{aligned} \tag{23}$$

The patient can be *hard-assigned* to each group by using the Bayes' rule: let κ_i denote to which group the patient has been assigned. According to this rule we obtain that $\kappa_i = h$ if $q_{ih} \geq q_{ij}$ for $j = 0, \ldots, K$, that is, the patient is assigned to the group for which $q_{i\cdot}$ is highest.

For this dataset we could observe for example that 234 out of 236 censored patients are classified as corresponding to the group with $H = 1$, which is characterized by a lower hazard function for the primary event (since $\gamma_1^T < 0$) and (obviously) a higher hazard for the censoring event. In addition, patients with lower baseline CD4 cell count are more likely to be classified in the group with $H = 0$ (mean 345.64 against a mean of 365.28 for $H = 1$). Further analysis with other covariates are likewise possible. We hereby take into account the CD4 as it is part of the hazard function specification for T and C.

Extensions

Analysis of competing risks

We can extend the approach described so far to the analysis of the joint distribution of $(T_1, \ldots, T_M)$ where T_j is the time to event for the jth cause of decrement ($j = 1, \ldots, M$). Within a competing risks framework only $T' = \min(\mathrm{T}_1, \ldots, \mathrm{T}_\mathrm{M})$ can be observed.

Suppose that conditional on the latent factor H, $(T_1, \ldots, T_M)$ is a vector of pairwise independently distributed times to event, whose distribution is characterized by the following hazard function:

$$\mu^m\big(t \mid x^m(t), h; \beta^m, \gamma^m\big) = \mu_0^m(t)\exp\big(\beta^{m'} x^m(t) + \gamma_1^m h_1 + \ldots + \gamma_K^m h_K\big) \tag{24}$$

for $m = 1, \ldots, M$

In this way, we can also account for further independence assumptions: for example, if $\gamma^m = (\gamma_1^m, \ldots, \gamma_K^m) = 0$, then T_m is independently distributed with respect to the vector $(T_1, \ldots, T_{m-1}, T_{m+1}, \ldots, T_M)$.

More generally, instead of having one common latent factor for all competing risks, we can have H_m $(m = 1, \ldots, M)$, such that the statistical association among time to events is characterized by the joint distribution of $(H_1, \ldots, H_M)$.

Clustered units

Suppose we deal with units divided into clusters, such as patients in different clinics, each corresponding to a level of the categorical variable G, whose value is known for each patient.

It is possible to account for the specific cluster either by including G as a factor in the hazard function, or by letting G to explain the heterogeneity in the distribution of H for each unit.

In the latter case, we can reasonably assume that G is independently distributed with respect to T and C conditional on x^T, x^C and H. Then, the joint distribution of (T, C) for the ith individual becomes:

$$\begin{aligned} & f\left(t, c \mid x_i^T(t), x_i^C(c); \beta^T, \gamma^T, \beta^C, \gamma^C\right) \\ & = \sum_{k=0}^{K} \mathrm{Pr}\left(\mathrm{H} = \mathrm{k} \mid \mathrm{g_i}\right)\left[\ \exp\left(-\int_0^{\mathrm{t}} \mu_0^{\mathrm{T}}(\mathrm{s}) \exp\left(\beta^{\mathrm{T}'} \mathrm{x_i^T}(\mathrm{s}) + \gamma_{\mathrm{k}}^{\mathrm{T}}\right)\mathrm{ds}\right)\right. \\ & \left. \times \mu_0^T(t) \exp\left(\beta^{T'} x_i^T(t) + \gamma_k^T\right)\right] \\ & \times \left[\exp\left(-\int_0^c \mu_0^C(s) \exp\left(\beta^{C'} x_i^C(s) + \gamma_k^C\right)\mathrm{d}s\right)\left(\mu_0^C(c) \exp\left(\beta^{C'} x_i^C(c) + \gamma_k^C\right)\right)\right] \end{aligned} \tag{25}$$

This modelling framework allows also to compare clusters. One possibility is the use of the Hellinger distance, which is used for example in topic models to qualitatively compare the topical content of two documents ([45]). Let CL1 and CL2 be any two different known clusters for the units in the sample: the Hellinger distance $d_H(\mathrm{CL1}, \mathrm{CL2})$ is

$$d_H(\mathrm{CL1}, \mathrm{CL2}) = \sum_{k=0}^{K}\left(\sqrt{\mathrm{Pr}\left(\mathrm{H} = \mathrm{k} \mid \mathrm{G} = \mathrm{CL1}\right)} - \sqrt{\mathrm{Pr}\left(\mathrm{H} = \mathrm{k} \mid \mathrm{G} = \mathrm{CL2}\right)}\right)^2 \tag{26}$$

When Bayesian techniques are used for inference, as we do in the present work, then the Hellinger distance is given by the expectation of $d_H(\cdot, \cdot)$ with respect to the posterior distribution of the parameters.

Conclusions

This work illustrated the potential of the use of latent factors to accomodate for the possibility of informative censoring in survival analysis. We rejoined existing literature, and described how we can generalize the specification of the heterogeneity component of the hazard function. We also emphasized how this methodology can be extended to the analysis of competing risks (Section 6.1) and of data from different known clusters (Section 6.2).

The inferential challenges of this type of ill-posed models have been addressed by means of a fully Bayesian approach and we described how modern computational tools such as Hamiltonian Monte Carlo methods are strongly recommended in these circumstances.

We applied those methodologies to the analysis of the ACTG175 clinical trial and we found that modelling the heterogeneity and the presence of informative censoring turns out to improve the model fit in terms of information criterion when compared with standard approaches such as the assumption of non-informative censoring.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

ORCID iD

Francesco Ungolo https://orcid.org/0000-0002-6642-0425

Notes

1. See [24] for an explanation of local and global identifiability.
2. Taking into account that $\gamma_0^T = \gamma_0^C = 0$.
3. The data can be downloaded as supplementary material of the book of [38]

References

1. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proc Natl Acad Sci U.S.A* 1975; **72**: 20–22.
2. Crowder M. On assessing independence of competing risks when failure times are discrete. *Lifetime Data Anal* 1996; **2**: 195–209.
3. Crowder M. A test for independence of competing risks with discrete failure times. *Lifetime Data Anal* 1997; **3**: 215.
4. Emoto SE and Matthews PC. A weibull model for dependent censoring. *Ann Statist* 1990; **18**: 1556–1577.
5. Zheng M and Klein JP. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* 1995; **82**: 127–138.
6. Scharfstein DO and Robins JM. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika* 2002; **89**: 617–634.
7. Jackson D, White IR, Seaman S, et al. Relaxing the independent censoring assumption in the cox proportional hazards model using multiple imputation. *Stat Med* 2014; **33**: 4681–4694.
8. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society Series B (Methodological)* 1972; **34**: 187–220.
9. Huang X and Wolfe RA. A frailty model for informative censoring. *Biometrics* 2002; **58**: 510–520.
10. Rowley M, Garmo H, Van Hemelrijck M, et al. A latent class model for competing risks. *Stat Med* 2017; **36**: 2100–2119.
11. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res* 2010; **11**: 3571–3594.
12. Drton M and Plummer M. A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2017; **79**: 323–380.
13. Pereyra M. Maximum-a-posteriori estimation with Bayesian confidence regions. *SIAM J Imaging Sci* 2017; **10**: 285–302.
14. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974; **19**: 716–723.
15. Schwarz G. Estimating the dimension of a model. *Ann Statist* 1978; **6**: 461–464.
16. Gelman A, Carlin JB, Stern HS, et al. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.
17. Hammer SM, Katzenstein DA, Hughes MD, et al. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *N Engl J Med* 1996; **335**: 1081–1090.
18. Lindsay BG. *Properties of the Maximum Likelihood Estimator of a Mixing Distribution*. Dordrecht: Springer Netherlands, 1981. ISBN 978-94-009-8552-0, 1981. pp. 95109. DOI: 10.1007/978-94-009-8552-0 8.
19. Lindsay BG. The geometry of mixture likelihoods: A general theory. *Ann Statist* 1983a; **11**: 86–94.
20. Lindsay BG. the geometry of mixture likelihoods, part II: The exponential family. *Ann Statist* 1983b; **11**: 783–792.
21. Heckman J and Singer B. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 1984; **52**: 271–320.
22. Heckman JJ and Honorè BE. The identifiability of the competing risks model. *Biometrika* 1989; **76**: 325–330.
23. Abbring JH and Van Den Berg GJ. The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2003; **65**: 701–710.
24. Catchpole EA and Morgan BJT. Detecting parameter redundancy. *Biometrika* 1997; **84**: 187–196.
25. McLachlan GJ and Peel D. *Finite mixture models*. Wiley Series in Probability and Statistics, New York, 2000.
26. Titterington DM, Smith AFM and Makov UE. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.
27. Marin J-M, Mengersen KL and Robert C. Bayesian modelling and inference on mixtures of distributions. In Dey, D. and Rao, C., editors, *Handbook of Statistics: Volume 25*. Elsevier, 2005.
28. Betancourt MJ. Identifying Bayesian Mixture Models, https://betanalpha.github.io/assets/case˙studies/identifying˙mixture˙models.html, 2017.
29. Stan Development Team, Stan reference manual. http://mc-stan.org/, version 2.18, 2018.
30. Watanabe S. A widely applicable Bayesian information criterion. *J Mach Learn Res* 2013; **14**: 867–897.
31. Neal RM. MCMC using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* 2010; **54**: 113–162.
32. Stan Development Team. RStan: the R interface to Stan. R package version 2.17.3, http://mc-stan.org/, 2018.
33. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/, 2013.
34. Hoffman MD and Gelman A. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian monte carlo. *J Mach Learn Res* 2014; **15**: 1593–1623.
35. Kalbfleisch JD. Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society Series B (Methodological)* 1978; **40**: 214–221.
36. Bremhorst V and Lambert P. Flexible estimation in cure survival models using Bayesian P-splines. *Comput Stat Data Anal* 2016; **93**: 270–284. DOI: 10.1016/j.csda.2014.05.009. http://www.sciencedirect.com/science/article/pii/S0167947314001492.

37. Ibrahim JG, Chen MH and Sinha D. *Bayesian Survival Analysis*, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer Verlag, 2001, pp. 479, ISBN: 0-387-95277-2, 2001.
38. Elashoff R, Li G and Li N. *Joint Modeling of Longitudinal and Time-to-Event Data*. Chapman and Hall/CRC, DOI:10.1201/9781315374871, 2016.
39. Rotnitzky A, Farall A, Bergesio A, et al. Analysis of failure time data under competing censoring mechanisms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2007; **69**: 307–327. DOI: 10.1111/j.1467-9868.2007.00590.x.
40. Muggeo VM. segmented: an R package to fit regression models with broken-line relationships. *R News* 2008; **8**: 20–25. https://cran.r-project.org/doc/Rnews/.
41. Gelman A and Rubin DB. Inference from iterative simulation using multiple sequences. *Statist Sci* 1992; **7**: 457–472. DOI: DOI:10.1214/ss/1177011136.
42. Moore CM, Carlson NE, MaWhinney S, et al. A dirichlet process mixture model for non-ignorable dropout. *Bayesian Analysis* 2020; **15**: 1139–1167. DOI: 10.1214/19-BA1181.
43. Cox DR and Snell EJ. A general definition of residuals. *Journal of the Royal Statistical Society Series B (Methodological)* 1968; **30**: 248–275.
44. Therneau TM. *A Package for Survival Analysis in R*. R package version 3.1–12, 2020. https://CRAN.R-project.org/package=survival.
45. Blei DM and Lafferty JD. A correlated topic model of science. *Ann Appl Stat* 2007; **1**: 17–35.

Appendix

Results of the simulation study for $K = 2$

Table A1. Posterior mean and coverage by 95% credible intervals (95% Cred. int.) across the simulations for the fitted models with $K = 0, 1, 2, 3$ with true $K = 2$ and $n = 500$.

Par.	True v.	Post. mean			**95**% Cred. int.				
		K=0	K=1	K=2	K=3	K=0	K=1	K=2	K=3
β_1^T	0.3	0.30	0.25	0.31	0.32	91	96	97	97
β_2^T	−1	−0.94	−0.92	−1.07	−1.09	89	94	98	96
γ_1^T	3	–	–	2.99	–	–	–	100	–
γ_2^T	2	–	–	−1.25	–	–	–	92	–
β_1^C	−0.7	−0.53	−0.60	−0.71	−0.75	77	90	97	97
β_2^C	0.5	0.38	0.41	0.51	0.53	76	92	96	98
γ_1^C	2.5	–	–	2.41	–	–	–	97	–
γ_2^C	4	–	–	4.06	–	–	–	97	–
ζ_0	0.48	–	–	0.47	–	–	–	93	–
ζ_1	0.32	–	–	0.28	–	–	–	100	–
ζ_2	0.20	–	–	0.18	–	–	–	100	–

Further results from the case study

Graphical posterior checks

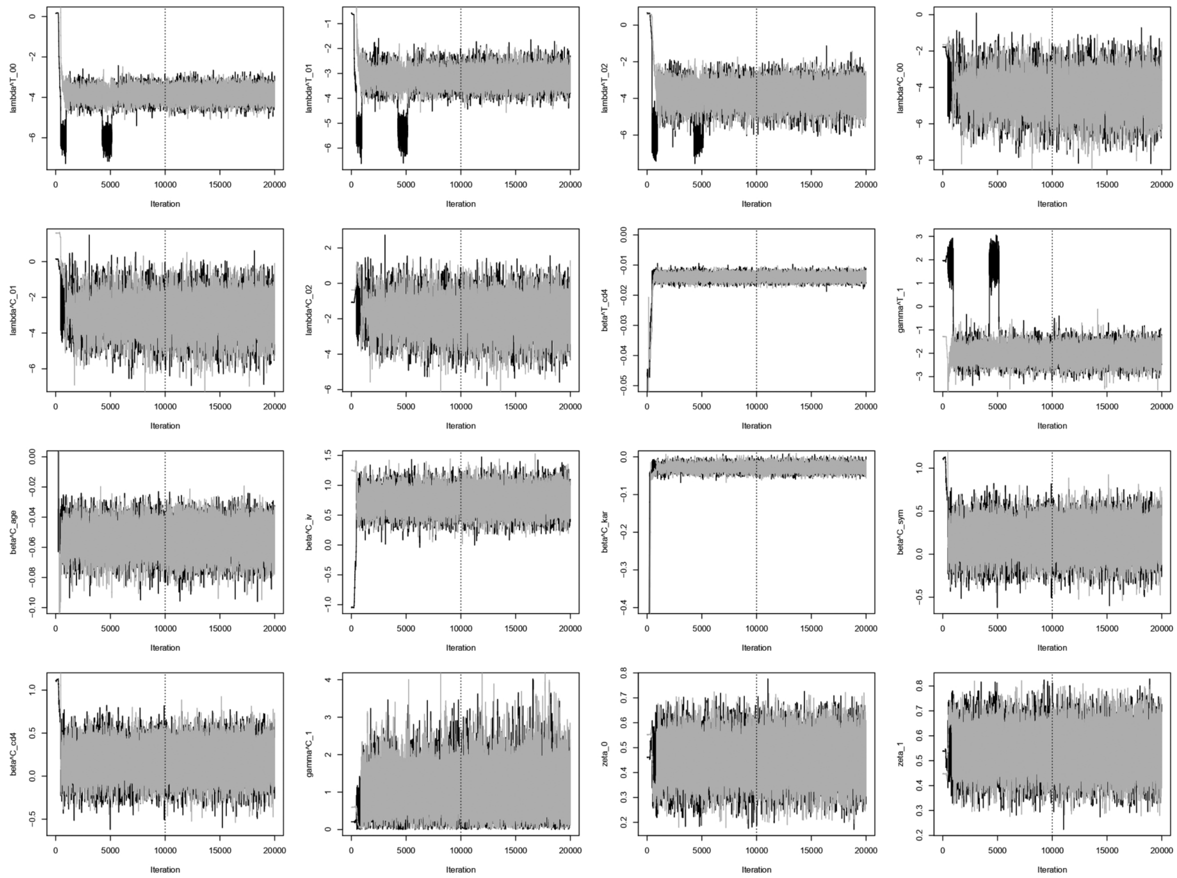

Figure A1. Traceplots of the first two chains for the model with $K = 1$. The vertical dotted line separates the first 10,000 warm up samples from the samples used to approximate the posterior distribution.

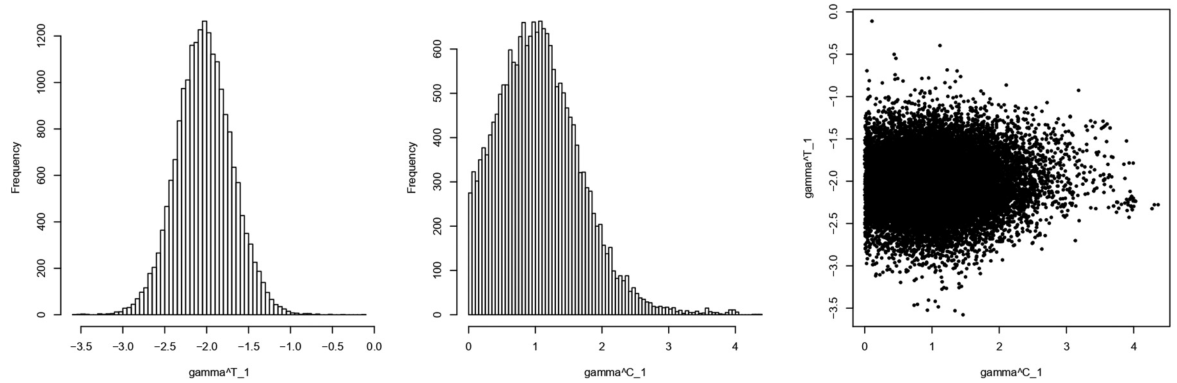

Figure A2. Histogram and scatterplot of the posterior samples of γ_1^T and γ_1^C.

Results from model with $K = 0, 2, 3$

Table A2. Summary of the posterior distribution of the parameters for the model $K = 0$.

Parameter	Mean	St. dev.	2.5-th%	50-th%	97.5-th%
λ^T_{00}	−5.18	0.20	−5.58	−5.18	−4.79
λ^T_{01}	−4.73	0.15	−5.03	−4.73	−4.43
λ^T_{02}	−5.33	0.46	−6.34	−5.31	−4.52
λ^C_{00}	−3.79	0.91	−5.60	−3.78	−2.04
λ^C_{01}	−2.46	0.91	−4.27	−2.45	−0.68
λ^C_{02}	−1.36	0.91	−3.15	−1.35	0.40
$\beta^T_{\mathrm{CD4(t)}}$	−0.0108	0.0007	−0.0125	−0.0111	−0.0097
β^C_{age}	−0.05	0.01	−0.07	−0.05	−0.03
β^C_{iv}	0.73	0.17	0.38	0.73	1.04
β^C_{kar}	−0.03	0.01	−0.04	−0.03	−0.01
β^C_{sym}	0.21	0.18	−0.15	0.21	0.55
$\beta^C_{\mathrm{CD4(t)}}$	−0.0013	0.0005	−0.0023	−0.0013	−0.00032

Table A3. Summary of the posterior distribution of the parameters for the model $K = 2$.

Parameter	Mean	St. dev.	2.5-th%	50-th%	97.5-th%
λ^T_{00}	−4.00	0.40	−4.84	−3.99	−3.26
λ^T_{01}	−3.32	0.41	−4.14	−3.32	−2.52
λ^T_{02}	−3.91	0.61	−5.19	−3.90	−2.75
λ^C_{00}	−4.12	1.08	−6.28	−4.11	−2.06
λ^C_{01}	−2.68	1.09	−4.87	−2.67	−0.58
λ^C_{02}	−1.46	1.10	−3.62	−1.44	0.66
$\beta^T_{\mathrm{CD4(t)}}$	−0.0141	0.00112	−0.0164	−0.014	−0.012
γ^T_1	−1.97	0.76	−2.98	−2.06	0.86
γ^T_2	−3.81	5.23	−18.23	−2.18	1.85
β^C_{age}	−0.06	0.01	−0.08	−0.06	−0.04
β^C_{iv}	0.84	0.21	0.44	0.84	1.27
β^C_{kar}	−0.03	0.01	−0.05	−0.03	−0.01
β^C_{sym}	0.21	0.21	−0.22	0.21	0.61
$\beta^C_{\mathrm{CD4(t)}}$	−0.0009	0.0005	−0.002	−0.00084	−0.00006
γ^C_1	0.87	0.54	0.06	0.81	2.02
γ^C_2	2.17	0.92	0.57	2.11	4.09
ζ_0	0.50	0.09	0.32	0.50	0.67
ζ_1	0.37	0.08	0.21	0.37	0.54
ζ_2	0.13	0.07	0.03	0.12	0.29

Table A4. Summary of the posterior distribution of the parameters for the model $K = 3$.

Parameter	Mean	St. dev.	2.5-th%	50-th%	97.5-th%
λ_{00}^{T}	−3.95	0.47	−4.92	−3.93	−3.07
λ_{01}^{T}	−3.24	0.48	−4.15	−3.24	−2.27
λ_{02}^{T}	−3.82	0.65	−5.10	−3.81	−2.54
λ_{00}^{C}	−4.13	1.13	−6.32	−4.13	−1.98
λ_{01}^{C}	−2.64	1.15	−4.86	−2.65	−0.41
λ_{02}^{C}	−1.38	1.16	−3.59	−1.40	0.91
$\beta_{\mathrm{CD4(t)}}^{T}$	−0.0147	0.0013	−0.0173	−0.0147	−0.0124
γ_{1}^{T}	−1.67	1.13	−3.21	−1.90	1.55
γ_{2}^{T}	−2.41	3.45	−12.33	−2.00	1.70
γ_{3}^{T}	−4.38	5.54	−18.83	−2.67	2.16
β_{age}^{C}	−0.06	0.01	−0.09	−0.06	−0.04
β_{iv}^{C}	0.88	0.23	0.46	0.87	1.34
β_{kar}^{C}	−0.03	0.01	−0.05	−0.03	−0.01
β_{sym}^{C}	0.21	0.21	−0.22	0.21	0.62
$\beta_{\mathrm{CD4(t)}}^{C}$	−0.00097	0.0006	0.0022	−0.0009	−0.00007
γ_{1}^{C}	0.79	0.56	0.04	0.70	2.09
γ_{2}^{C}	1.52	0.68	0.35	1.48	2.99
γ_{3}^{C}	2.92	1.21	1.09	2.76	5.75
ζ_{0}	0.42	0.09	0.24	0.42	0.60
ζ_{1}	0.29	0.08	0.14	0.29	0.47
ζ_{2}	0.21	0.08	0.06	0.21	0.38
ζ_{3}	0.08	0.05	0.01	0.07	0.22