Bayesian semi parametric multi-state models

Thomas Kneib and Andrea Hennerfeind

Statistical Modelling 2008 8: 169
DOI: 10.1177/1471082X0800800203

Statistical Modelling 2008; **8**(2): 169–198

Bayesian semiparametric multi-state models

Thomas Kneib[1] and Andrea Hennerfeind[1]
[1] Department of Statistics, Ludwig-Maximilians-University, Germany

Abstract: Multi-state models provide a unified framework for the description of the evolution of discrete phenomena in continuous time. One particular example is Markov processes which can be characterised by a set of time-constant transition intensities between the states. In this paper, we will extend such parametric approaches to semiparametric models with flexible transition intensities based on Bayesian versions of penalised splines. The transition intensities will be modelled as smooth functions of time and can further be related to parametric as well as nonparametric covariate effects. Covariates with time-varying effects and frailty terms can be included in addition. Inference will be conducted either fully Bayesian (using Markov chain Monte Carlo simulation techniques) or empirically Bayesian (based on a mixed model representation). A counting process representation of semiparametric multi-state models provides the likelihood formula and also forms the basis for model validation via martingale residual processes. As an application, we will consider human sleep data with a discrete set of sleep states such as REM and non-REM phases. In this case, simple parametric approaches are inappropriate since the dynamics underlying human sleep are strongly varying throughout the night and individual-specific variation has to be accounted for using covariate information and frailty terms.

Key words: frailties; martingale residuals; multi-state models; penalised splines; time-varying effects; transition intensities

Received November 2006; **revised** May 2007; **accepted** July 2007

1 Introduction

Multi-state models are a flexible tool for the analysis of time-continuous phenomena that can be described by a discrete set of states. Such data structures naturally arise when observing a discrete response variable for several individuals or objects continuously over time. Some common examples are depicted in Figure 1 in terms of their reachability graphs for illustration. For recurrent events (Figure 1 (a)), the individual observations evolve through time, moving repeatedly between a fixed set of states. Our application on sleep research will be of this type, where the states are given by the sleep states awake, REM and Non-REM; compare also Figure 2 which shows two exemplary realisations of such sleep processes. Other model classes involve

Address for correspondence: Thomas Kneib, Department of Statistics, Ludwig-Maximilians-University, D-80539 Munich, Germany. E-mail: thomas.kneib@stat.uni-muenchen.de

10.1177/1471082X0800800203

absorbing states, for example, disease progression models (Figure 1 (b)), that are used to describe the chronological development of a certain disease. If the severity of this disease can be grouped into $Q-1$ ordered stages of increasing severity, a reasonable model might look like this: Starting from disease state 'q', an individual can only move to contiguous states, that is, either the disease gets worse and the individual moves to state '$q+1$', or the disease attenuates and the individual moves to state '$q-1$'. In addition, death is included as a further, absorbing state 'Q', which can be reached from any of the disease states. A model with several absorbing states is the competing risks model (Figure 1 (c)) where, for example, different causes of death are analysed simultaneously.

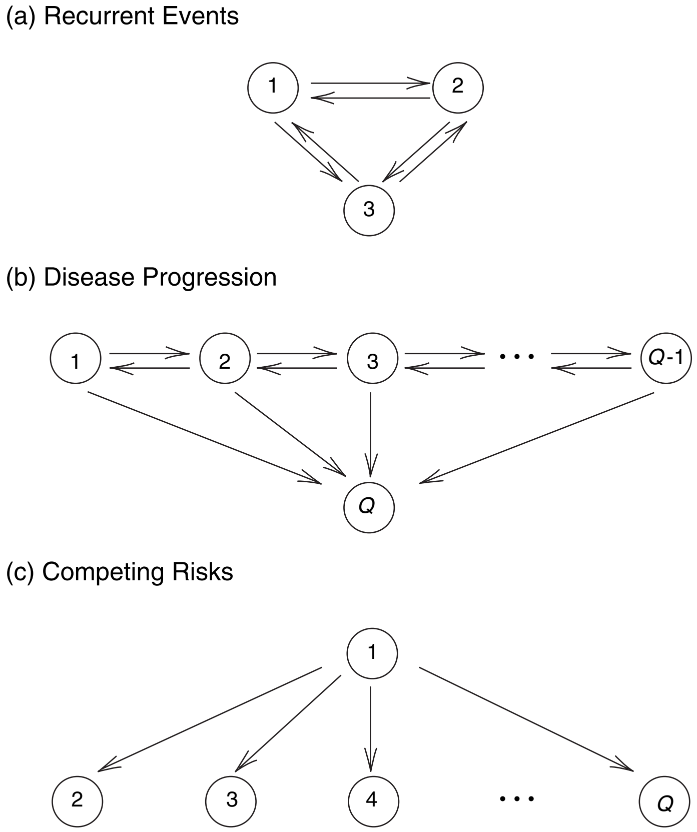

Figure 1 Reachability graphs of some common multi-state models

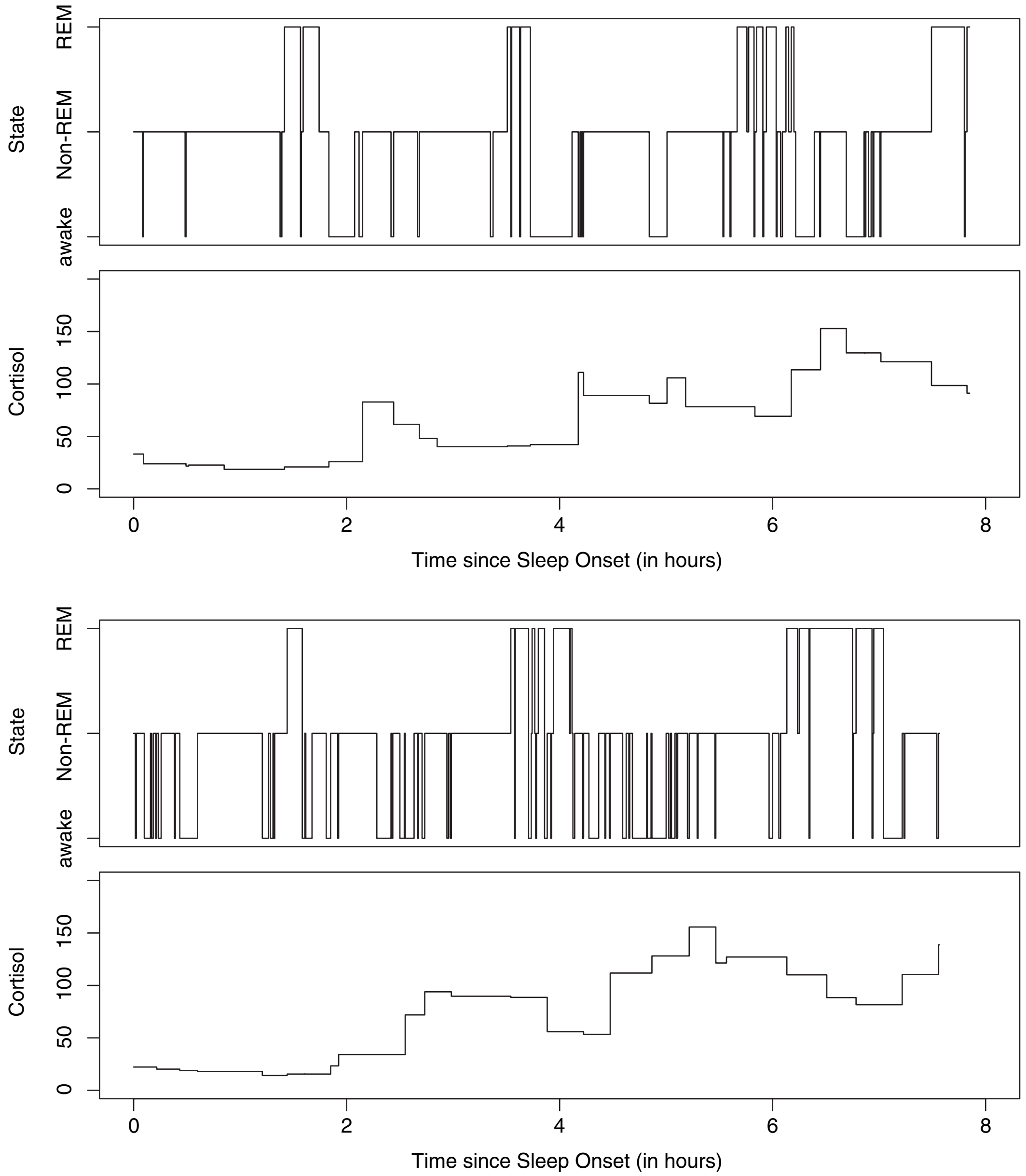

Figure 2 Realisations of two individual sleep processes and corresponding nocturnal cortisol secretion

As Figure 1 suggests, multi-state models can be described in terms of transitions between the states. The most simple model of this type are discrete Markov processes, where each of the transitions is associated with one time-constant transition intensity $\lambda^{(h)} > 0$, with $h = 1, \ldots, H$, indexing the set of possible transitions. Of course, such a purely parametric model is only appropriate if the dynamics underlying the transitions do not change over time. This assumption, however, does not apply to numerous real data applications. For example, in our application we will analyse human sleep data. It is well known that the dynamics of human sleep are strongly changing throughout the night with, for example, an increased propensity to switch to the REM state at the end of the night.

More flexible multi-state models have been introduced within two different frameworks: Aalen *et al.* (2004) considered dynamic versions of multi-state models based on Aalen‘s additive risk model. Such models rely heavily on the embracing framework of counting processes, (compare Andersen *et al.* 1993), and estimation is based on martingale theory. Fahrmeir and Klinger (1998) and Yassouridis *et al.* (1999) modelled the transition intensities in a Cox-type manner with smoothing splines for time-varying effects. In their approach, estimation is based on a backfitting scheme with internal smoothing parameter selection by AIC optimisation.

In this article, we extend the ideas by Fahrmeir and Klinger (1998) and propose a general semiparametric class of multi-state models that comprises the following features:

- Flexible modelling of baseline transition intensities in terms of penalised splines,
- Inclusion of parametric, time-varying and nonparametric covariate effects,
- Inclusion of frailty terms (that is, subject-specific random effects) to account for unobserved heterogeneity.

Estimation is based on a unified Bayesian formulation that incorporates penalised splines and random effects into one general framework. Inference can be conducted either fully Bayesian, based on Markov chain Monte Carlo (MCMC) simulation techniques, or empirically Bayesian, based on a mixed model representation. Both inferential procedures borrow from the time-continuous duration time models presented in Hennerfeind *et al.* (2006) and Kneib and Fahrmeir (2007), and allow for the simultaneous determination of all effects and smoothing parameters. Implementations are available in the free software package BayesX (visit http://www.stat.uni-muenchen.de/~bayesx for further information).

As an illustration of our approach, we will analyse data on human sleep collected at the Max-Planck Institute for Psychiatry in Munich as part of a larger study on sleep withdrawal. Our major concern is to obtain a valid description of the sleeping process of healthy participants of the study while accounting for possible covariate effects (for example, nocturnal hormonal secretion) and patient-specific individual sleeping habits. The performance of the developed models is assessed using martingale residuals and compared to parametric Markov process models. The data and code for reproducing our results will be made available in the Statistical Modelling Archives (see http://stat.uibk.ac.at/SMIJ/).

The structure of the paper is as follows: Section 2 describes the specification of hazard rates for the transitions and the corresponding prior assumptions. In Section 3 we introduce a counting process representation of the model that provides us with the likelihood formula for multi-state models. In addition, martingale residual processes can be derived from counting process theory, forming the basis for some of the model validation tools considered in Section 4. Section 5 outlines inferential schemes while Sections 6 and 7 contain the results of our application and of a small simulation study, respectively. The concluding Section 8 comments on directions of future research.

2 Specification of multi-state models in terms of hazard rates

A multi-state model is fully described by a set of (possibly individual-specific) hazard rates $\lambda_i^{(h)}(t)$ where h, $h = 1, \dots, H$, indexes the type of the transition and i, $i = 1, \dots, n$, indexes the individuals. Since the hazard rates describe durations between transitions, we specify them in analogy to hazard rate models for continuous time survival analysis. To be more specific, $\lambda_i^{(h)}(t)$ is modelled in a multiplicative Cox-type way as

$$\lambda_i^{(h)}(t) = \exp(\eta_i^{(h)}(t)),$$

where

$$\eta_i^{(h)}(t) = g_0^{(h)}(t) + \sum_{l=1}^{L} g_l^{(h)}(t) u_{il}(t) + \sum_{k=1}^{K} f_k^{(h)}(x_{ik}(t)) + v_i(t)'\gamma^{(h)} + b_i^{(h)} \qquad (2.1)$$

is an additive predictor consisting of the following components:

- A time-varying, nonparametric baseline effect $g_0^{(h)}(t)$ common for all observations.
- Covariates $u_{il}(t)$ with time-varying effects $g_l^{(h)}(t)$. In our application, $u_{il}(t)$ will represent the current level of a certain hormone; hence $u_{il}(t)$ by itself is time-varying but its effect is also varying throughout the night.
- Nonparametric effects $f_k^{(h)}(x_{ik}(t))$ of continuous covariates $x_{ik}(t)$. For example, we might also include the hormonal level in a nonparametric way.
- Parametric effects $\gamma^{(h)}$ of covariates $v_i(t)$.
- Frailty terms $b_i^{(h)}$ to account for unobserved heterogeneity.

For each individual, we assume that a full transition path is observed consisting of the following parts: A set of ordered time points $0 = S_{i0} < S_{i1} < \dots < S_{ir}$

$< \ldots < S_{im_i}$ indicating the time points of transitions for individual i, the states $Y_i(S_{ir})$, $r = 0, \ldots, m_i$, the individual moves to at the end of a time interval, and indicators $\delta_i^{(h)}(t)$ for type h transitions at time t. From the time points of transitions, we can deduce the duration time in the rth state as $T_{ir} = S_{ir} - S_{i,r-1}$, $r = 1, \ldots, m_i$.

After reindexing, the predictor vectors $\eta^{(h)} = (\eta_1^{(h)}(S_{1,0}), \ldots, \eta_1^{(h)}(S_{1,m_1}), \ldots, \eta_n^{(h)}(S_{n,0}), \ldots, \eta_n^{(h)}(S_{n,m_n})'$ can be represented in generic notation as

$$\eta^{(h)} = V_1^{(h)}\xi_1^{(h)} + \ldots + V_J^{(h)}\xi_J^{(h)} + V^{(h)}\gamma^{(h)}, \tag{2.2}$$

where $V^{(h)}$ corresponds to the usual design matrix of fixed effects. The construction of the design matrices $V_1^{(h)}, \ldots, V_J^{(h)}$ for time-varying, nonparametric and random effects will be described in the following discussion of prior assumptions.

To model time-varying and nonparametric effects, we employ penalised splines, a parsimonious yet flexible approach to represent smooth functions. For the sake of simplicity, we will drop the transition and the covariate index in the following discussion. The basic idea of penalised splines (Eilers and Marx, 1996) is to represent a function $f(x)$ (or $g(t)$) of a smooth covariate x (or of time t) as a linear combination of a large number of B-spline basis functions, that is,

$$f(x) = \sum_{m=1}^{M} \xi_m B_m(x).$$

Instead of estimating the resulting regression coefficients $\xi = (\xi_1, \ldots, \xi_M)'$ unrestricted, a penalty term is added to the likelihood to enforce smoothness of the estimated function. From a Bayesian perspective, this corresponds to a smoothness prior for ξ (Brezger and Lang, 2006). Since the derivatives of B-splines are determined by the magnitude of the differences in adjacent parameter values, a sensible prior distribution can be obtained by assuming a Gaussian distribution with appropriate variance for these differences. This corresponds to a random walk prior for the sequence of regression coefficients, that is,

$$\xi_m = \xi_{m-1} + \varepsilon_m, \quad m = 2, \ldots, M, \tag{2.3}$$

for a first order random walk or

$$\xi_m = 2\xi_{m-1} - \xi_{m-2} + \varepsilon_m, \quad m = 3, \ldots, M, \tag{2.4}$$

for a second order random walk and Gaussian error terms $\varepsilon_m \sim N(0, \tau^2)$. In addition, noninformative, flat priors are assigned to the initial values. The variance parameter of the error term can now be interpreted analogously to a smoothing parameter. For large variances, the random walk prior allows for ample deviations in the differences of adjacent parameters while a small variance enforces smaller differences and, as a consequence, smoother function estimates are obtained.

In vector-matrix notation, penalised splines lead to the following representation for the function evaluations defining predictor (2.1): The baseline hazard rate can be expressed as $g_0^{(h)}(t) = v_0^{(h)}(t)'\xi_0^{(h)}$ where $v_0^{(h)}(t) = (B_{01}^{(h)}(t), \ldots, B_{0M}^{(h)}(t))'$ and $\xi_0^{(h)} = (\xi_{01}^{(h)}, \ldots, \xi_{0M}^{(h)})'$. Similarly, we obtain $g_l^{(h)}(t)u_l(t) = v_l^{(h)}(t)'\xi_l^{(h)}$ for the time-varying effects with $v_l^{(h)}(t) = (u_l(t)B_{l1}^{(h)}(t), \ldots, u_l(t)B_{lM}^{(h)}(t))'$ and $f_k^{(h)}(x_k(t)) = v_k^{(h)}(t)'\xi_k^{(h)}$ for nonparametric effects with $v_k^{(h)}(t) = (B_{k1}^{(h)}(x_k(t)), \ldots, B_{kM}^{(h)}(x_k(t)))'$. The design matrices in (2.2) are then obtained by stacking the design vectors. In all cases, the vectors of regression coefficients follow a multivariate Gaussian prior derived from the random walk assumptions. The density of these distributions can be expressed as

$$p(\xi|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\xi' P \xi\right) \tag{2.5}$$

where the precision matrix $P = D'D$ is defined by the crossproduct of appropriate difference matrices D. Note that in general, distribution (2.5) is improper since P does not have full rank due to the improper distributions of the initial values in the random walk definition.

To complete the Bayesian model formulation, we assign noninformative, flat priors to the fixed effects, that is, $p(\gamma^{(h)}) \propto const$, and i.i.d. Gaussian priors $b_i^{(h)} \sim N(0, \tau_h^2)$ with transition-specific variances to the frailty terms. Note that these random effects distributions can also be cast into the multivariate form (2.5) by simply collecting all the random effects for one transition in the vector $\xi = (b_1^{(h)}, \ldots, b_n^{(h)})'$ and defining $P = I_n$. The design matrix for random effects is given by a 0/1-incidence matrix which ties together a specific individual and its random effect. Note that the possibility to cast both random effects and penalised splines into one general framework considerably facilitates implementation of inferential procedures since the same algorithms can be used for both penalised splines and random effects.

Finally, for the variance parameters τ^2 determining the variability of either nonparametric function estimates or random effects, we will consider two situations: In the first case, the variances are treated as fixed unknown constants, that are to be estimated from their marginal posterior. This corresponds to empirical Bayes estimation and will be further discussed in Section 5.1. In the second case, additional inverse gamma-type hyperpriors are assigned to the variances. This corresponds to a fully Bayesian approach and will be described in Section 5.2.

3 Counting process representation and likelihood contributions

For each individual i, $i = 1, \dots, n$, the likelihood contribution in a multi-state model can be derived from a counting process representation of the multi-state model. Let $N_i^{(h)}(t)$, $h = 1, \dots, H$, be a set of counting processes, counting transitions of type h for individual i. Consequently, $h = 1, \dots, H$, indexes the observable transitions in the model under consideration and the jumps of the counting processes $N_i^{(h)}(t)$ are defined by the transition times S_{ir}, $r = 1, \dots, m_i$ of the corresponding multi-state process for individual i.

From the classical counting process theory (see for example, Andersen *et al.*, 1993, Ch. VII.2), the intensity processes $\alpha_i^{(h)}(t)$ of the counting processes $N_i^{(h)}(t)$ are derived as the product of the hazard rate for type h transitions $\lambda_i^{(h)}(t)$ and a predictable at-risk indicator process $I_i^{(h)}(t)$ (a sufficient condition for $I_i^{(h)}(t)$ to be predictable is that the paths are continuous from the left), that is,

$$\alpha_i^{(h)}(t) = I_i^{(h)}(t)\lambda_i^{(h)}(t),$$

where the hazard rates are constructed in terms of covariates as described in Section 2. The at-risk indicator $I_i^{(h)}(t)$ takes the value one if individual i is at risk for a type h transition immediately before time t, and zero otherwise. For example, in the multi-state model of Figure 1(a), an individual in state 2 is at risk for both transitions to state 1 and state 3. Hence, the at-risk indicators for both the transitions '2 to 1' and '2 to 3' will be equal to one as long as the individual remains in state 2.

Under mild regularity conditions, the individual log-likelihood contributions can now be obtained from counting process theory as

$$l_i = \sum_{h=1}^{H} \left[\int_0^{S_{i,m_i}} \log(\lambda_i^{(h)}(t)) dN_i^{(h)}(t) - \int_0^{S_{i,m_i}} \lambda_i^{(h)}(t) I_i^{(h)}(t) dt \right], \qquad (3.1)$$

where S_{i,m_i} denotes the time until which individual i has been observed. The likelihood contributions can be interpreted similarly as with hazard rate models for survival times (and in fact coincide with these in the case of a multi-state process with only one transition to an absorbing state). The first term corresponds to contributions at the transition times since the integral with respect to the counting process in fact equals a simple sum over the transition times. Each of the summands is then given by the log-intensity for the observed transition evaluated at this particular time point. In survival models, this term simply equals the log-hazard evaluated at the survival time for uncensored observations. The second term reflects cumulative intensities integrated over accordant waiting periods between two successive transitions. The integral is evaluated for all transitions the corresponding person is at risk at during the current period. In survival models, there is only one such transition (the transition from 'alive' to 'dead') and the integral is evaluated from the time of entrance to the study to the survival or censoring time.

These considerations yield an alternative representation of the likelihood, where each of the individual contributions is expressed in terms of the transition indicators $\delta_i^{(h)}(t)$ and observed transition times S_{ir}, $r = 0, \ldots, m_i$. The indicators $\delta_i^{(h)}(t)$ take the value one if a transition of type h is observed at time t for individual i, and zero otherwise, while the S_{ir} are defined by the times at which the corresponding individual experiences a transition. This leads to the alternative log-likelihood formula

$$l_i = \sum_{r=1}^{m_i} \sum_{h=1}^{H} \left[\delta_i^{(h)}(S_{ir}) \log(\lambda_i^{(h)}(S_{ir})) - I_i^{(h)}(S_{ir}) \int_{S_{i,r-1}}^{S_{ir}} \lambda_i^{(h)}(t)dt \right], \qquad (3.2)$$

which reveals more clearly the connection to the commonly known likelihood of hazard rate models in case of continuous survival times.

Under the usual assumption of conditional independence, the complete log-likelihood is given by the sum of the individual contributions. Note that the first integral in (3.1) reduces to a sum as shown in Equation (3.2) while the second integral has to be evaluated. When using splines of degree zero or one, explicit formulae for the integral can be derived. In general, however, some numerical integration technique has to be applied. In our implementation, we utilise the trapezoidal rule due to its simplicity but, of course, more sophisticated methods could also be used if required.

4 Model Validation

The counting process formulation of multi-state models also provides a possibility for model checking based on martingale residuals (compare Aalen *et al.* (2004) for a similar approach in the additive risk model). Since every counting process is a submartingale by construction, we can apply the Doob–Meyer decomposition (Andersen *et al.*, 1993, Chapter II.3) to $N_i^{(h)}(t)$ and obtain

$$\begin{aligned} N_i^{(h)}(t) &= A_i^{(h)}(t) + M_i^{(h)}(t) \\ &= \int_0^t \alpha_i^{(h)}(u)du + M_i^{(h)}(t), \end{aligned}$$

where $M_i^{(h)}(t)$ is a martingale and $A_i^{(h)}(t)$ is a predictable process called the compensator of $N_i^{(h)}(t)$. The compensator can be represented as the integral over the intensity process and is therefore also called the cumulative intensity process. The Doob–Meyer decomposition can be interpreted analogously to the decomposition of a times series into a trend (the compensator) and an error component (the martingale). Hence, replacing the compensator process with an estimate $\hat{A}_i^{(h)}(t)$ obtained from the model under consideration, yields estimated residual processes $N_i^{(h)}(t) - \hat{A}_i^{(h)}(t)$. If the model is valid, the estimated residuals should (approximately) have martingale properties. For example, their expectation should be zero and increments in non-overlapping intervals should be uncorrelated (Hall and Heyde, 1980, Sec. 1.6).

In addition to computing the residual processes for the estimation data, out-of-sample validation is also a useful tool that naturally avoids the risk of overfitting the data. Besides looking at residual paths in the validation data, it is also possible to compare predicted transitions to the actually observed ones. Given the sequence of transition times S_{ir}, the likelihood for a transition of type h at time S_{ir}, according to the estimated model, is given by

$$p_i^{(h)}(S_{ir}) = \frac{\hat{\alpha}_i^{(h)}(S_{ir})}{\sum_{h'=1}^{H} \hat{\alpha}_i^{(h')}(S_{ir})}.$$

Therefore, the most likely transition at time S_{ir} is the transition with maximum intensity process at this time. Based on the sequence of most likely transitions, we propose to compare the actually observed path with the predicted one and to summarise the deviation in terms of some misclassification measure.

A more direct approach for model checking would be to test, for example, smooth effects of some covariates against linear alternatives. This idea is particularly attractive in our model, since such a test can be based on testing a single parameter, namely the variance component of a smooth function. When using second order random walk priors for a nonparametric effect, the limiting case $\tau^2 \to 0$ yields exactly the linear model. Hence, it might be tempting to test the alternatives

$$H_0 : \tau^2 > 0 \quad \text{vs.} \quad H_1 : \tau^2 = 0$$

using, for example, a likelihood ratio test. However, the parameter τ^2 is on the boundary of the parameter space under the null hypothesis and, as a consequence, standard asymptotics do no longer apply. Although great efforts have been spent to extend the likelihood ratio theory in this direction (compare for example, Crainiceanu *et al.* (2005) or Greven *et al.* (forthcoming)), the methodology currently available is not readily applicable in our model class.

The Bayesian analogon to the aforementioned likelihood ratio test would be to modify the inverse Gamma prior of τ^2 to a mixture of a point mass in zero and the original inverse Gamma distribution. This allows for a positive a posteriori probability of a zero variance corresponding to a linear effect of the respective covariate. However, when naively implementing this approach, mixing problems are usually encountered due to the dependency between the sampled (non-negative) values of τ^2 and the point mass in zero. These problems can be solved in normal models by marginalisation (compare for example, Smith and Kohn (2002) or Früehwirth-Schnatter & Tüchler (forthcoming)) but are more difficult to handle in non-Gaussian models. Therefore, model validation in the application in Section 6 will be restricted to the consideration of martingale residuals.

5 Bayesian Inference

Based on the likelihood introduced in Section 3, we are now prepared to discuss Bayesian inference in multi-state models. In the following, we will differentiate between two perspectives on the estimation problem: In an empirical Bayes approach, the variance parameters of the smoothness priors (2.5) will be treated as unknown constants which are to be estimated from their marginal posterior. This will be facilitated by a mixed model representation of the predictor defining the transition hazards; see Section 5.1. In a fully Bayesian treatment of multi-state models, all parameters, including the variances, will be treated as random and estimated simultaneously using MCMC simulation techniques; see Section 5.2. Both inferential procedures borrow from approaches that have been recently developed for continuous time survival models (compare Kneib and Fahrmeir (2007) for the empirical Bayes version and Hennerfeind *et al.* (2006) for the fully Bayesian approach) and extend them to the more general setup of multi-state models.

5.1 Empirical Bayes inference

In an empirical Bayes approach, we differentiate between parameters of primary interest (the regression coefficients in our model) and hyperparameters (the variance parameters). While prior distributions are assigned to the former, the latter are treated as unknown constants which are to be estimated by maximising their marginal posterior. Plugging these estimates into the posterior and maximising the resulting expression with respect to the regression coefficients then yields posterior mode estimates (as compared to the empirical mean estimates obtained from MCMC simulation averages).

Empirical Bayes estimation in semiparametric regression models has been considerably facilitated by the insight that regression models with smoothness priors of the form (2.5) can be represented as mixed models with i.i.d. random effects (compare, for example, Fahrmeir *et al.* (2004) or Ruppert *et al.* (2003)). This representation has the advantage that partially improper priors can be split into an improper and a proper part, therefore enabling the application of mixed model methodology for estimation of the variance parameters.

To be more specific, let ξ be the vector of regression coefficients describing a model term with $r = \text{rank}(P) \leq \dim(\xi) = d$, where P is the prior precision matrix corresponding to ξ. Our aim is to express ξ in terms of an r-dimensional vector of random effects b and a $(d - r)$-dimensional vector of fixed effects β. This can be achieved by applying the decomposition

$$\xi = \tilde{X}\beta + \tilde{Z}b \tag{5.1}$$

with suitably chosen design matrices $\tilde{X}$ and $\tilde{Z}$ of dimensions $(d \times d - r)$ and $(d \times r)$, respectively. The following conditions are assumed for the transformation in (5.1):

(i) The compound matrix $(\tilde{X}\,\tilde{Z})$ has full rank to make (5.1) a one-to-one transformation.

(ii) $\tilde{X}'P = 0$ yielding a flat prior for β, that is, β can be interpreted as a vector of fixed effects.

(iii) $\tilde{Z}'P\tilde{Z} = I_r$ yielding an i.i.d. Gaussian prior for b, that is, $b \sim N(0, \tau^2 I_r)$ can be interpreted as a vector of i.i.d. random effects with variance τ^2.

Correspondingly the vector of function evaluations transforms to

$$V\xi = V(\tilde{X}\beta + \tilde{Z}b) = X\beta + Zb$$

with $X = V\tilde{X}$ and $Z = V\tilde{Z}$. Applying this decomposition to all nonparametric effects in the model leads to a variance components mixed model representation for each of the transition intensities. Note that additional identifiability restrictions have to be imposed on the reparametrisation to obtain a valid model formulation; also compare the discussion in Kneib and Fahrmeir (2007). Each of the nonparametric effects in a transition intensity yields a column of ones in the design matrix X which models the overall level of the corresponding function. To obtain a valid model specification, we include an intercept in each of the transition intensity models and delete the superfluous columns from the reparameterisation. This has a similar effect as imposing centering restrictions on nonparametric functions, which is a common strategy to obtain identifiable additive models. Note that we do not have to impose centering restrictions on time-varying effects $g_l(t)$.

We will now briefly outline mixed model based estimation of multi-state models. Since each of the transition intensities can, in fact, be considered a hazard rate in a time-continuous duration time model, we will not discuss every step in full detail but refer to the complete description in Kneib and Fahrmeir (2007).

In mixed model formulation, the log-posterior for all parameters is given by

$$l_p(\beta, b, \tau^2) = \sum_{i=1}^{n} l_i - \sum_{h=1}^{H}\sum_{j=1}^{J} \frac{1}{2\tau_{hj}^2} b_{hj}' b_{hj}, \tag{5.2}$$

where β, b and τ^2 are vectors collecting all fixed effects, random effects and variances, respectively. The first term in (5.2) corresponds to the sum of likelihood contributions obtained from Equation (3.1) while the second term consists of sums over all prior distributions in the model. Since (5.2) has the form of a penalised likelihood, the regression coefficients can be obtained as penalised maximum likelihood estimates. This corresponds to the determination of posterior mode estimates for given variance parameters. Actual maximisation can be achieved by a Newton–Raphson-type algorithm that extends the algorithm presented in Kneib and Fahrmeir (2007).

The variances themselves are to be obtained from the marginal posterior, that is, by maximising (5.2) after integrating out all regression coefficients:

$$l_{marg}(\tau^2) = \int l_p(\beta, b, \tau^2) d\beta db \rightarrow \max_{\tau^2}.$$

Of course, this integral can hardly be solved analytically or numerically in practice, since β and b will typically be high-dimensional. Therefore, we apply a Laplace approximation to $l_{marg}(\tau^2)$, similar in spirit to the approach in Breslow and Clayton (1993), yielding an approximate solution to the integral depending on current estimates $\hat{\beta}$ and $\hat{b}$. Computing the score function and expected Fisher information of the approximate marginal posterior allows to devise a Fisher-scoring scheme for the estimation of τ^2. Since now the estimation scheme of the regression coefficients depends on the variances and vice versa, we update both quantities in turn until convergence is reached.

5.2 Fully Bayesian inference

In contrast to the empirical Bayes approach, a fully Bayesian approach is based upon the assumption that both the parameters of primary interest and the hyperparameters (the variance parameters) are random. Prior distributions are not only assigned to the former but, in a further stage of the hierarchy, also to the latter. We routinely assign inverse Gamma priors $IG(a; b)$

$$p(\tau^2) \propto \frac{1}{(\tau^2)^{a+1}} \exp\left(-\frac{b}{\tau^2}\right) \tag{5.3}$$

to all variances. They are proper for $a > 0$, $b > 0$, and we use $a = b = 0.001$ as a standard choice for a weakly informative prior. Note that uniform priors for the variance ($a = -1$, $b = 0$) or the standard deviation ($a = -0.5$, $b = 0$) are special (improper) cases of prior (5.3), still leading to proper posteriors under regularity assumptions (see Fahrmeir and Kneib (forthcoming) for a detailed discussion). The Bayesian model specification is completed by assuming that all priors for parameters are (conditionally) independent.

Again, since each of the transition intensities can be considered a hazard rate in a time-continuous duration time model, we will only briefly comment on fully

Bayesian inference for multi-state models and refer to Hennerfeind *et al.* (2006) for more details. Let ξ denote the vector of all regression coefficients (in the original parameterisation) and τ^2 the vector of all variance parameters. Fully Bayesian inference is based on the entire posterior distribution

$$p(\xi, \tau^2 \mid data) \propto L(\xi, \tau^2)\, p(\xi, \tau^2),$$

where L denotes the likelihood (given by the product of the individual likelihood contributions) and $p(\xi, \tau^2)$ denotes the joint prior, which may be factorised due to the (conditional) independence assumption. Since the full posterior distribution is numerically intractable, we employ an MCMC simulation method that is based on updating full conditionals of single parameters or blocks of parameters (each with parameters corresponding to the same transition rate $\lambda_i^{(h)}(t)$), given the rest of the parameters and the data. Convergence of the Markov chains to their stationary distributions is assessed by inspecting the sampling paths and autocorrelation functions of the sampled parameters, which are used to estimate characteristics of the posterior distribution like means and standard deviations via their empirical analogues.

For updating the parameter vectors corresponding to time-independent functions $f_k^{(h)}$, as well as fixed effects $\gamma^{(h)}$ and frailty terms $b^{(h)}$, we use a slightly modified version of the Metropolis–Hastings algorithm based on iteratively weighted least squares (IWLS) proposals, developed for fixed and random effects in generalised linear mixed models by Gamerman (1997) and adapted to generalised additive mixed models in Brezger and Lang (2006). Suppose we want to update a certain parameter vector ξ, with current value ξ^c of the chain. Then a new value ξ^p is proposed by drawing a random vector from a (high-dimensional) multivariate Gaussian proposal distribution $q(\xi^c, \xi^p)$, which is obtained from a quadratic approximation to the posterior by a second order Taylor expansion with respect to ξ^c, in analogy to IWLS iterations in generalised linear models. More precisely, the goal is to approximate the posterior by a Gaussian distribution, obtained by accomplishing one IWLS step in every iteration of the sampler. Then, random samples have to be drawn from a high dimensional multivariate Gaussian distribution with precision matrix and mean

$$\tilde{P} = V'W(\xi^c)V + \frac{1}{\tau^2}P, \quad \tilde{m} = \tilde{P}^{-1}V'W(\xi^c)(\tilde{y} - \tilde{\eta}).$$

Here, $\tilde{\eta} = \eta - V\xi$ is the part of the linear predictor for the transition corresponding to ξ associated with all remaining effects, and $W(\xi^c) = \mathrm{diag}(w_{11}, \ldots, w_{1m_1}, \ldots, w_{nm_n})$ is the weight matrix for IWLS with weights calculated from the current state ξ^c as $w_{ir} = \int_{S_{i,r-1}}^{S_{ir}} I_i(u)\lambda_i(u)du$ for $r = 1, \ldots, m_i$, $i = 1, \ldots, n$. The vector of working observations $\tilde{y}$ is given by

$$\tilde{y} = W^{-1}(\xi^c)\Delta - \mathbb{1} + \eta$$

with $\Delta = \big(\delta_1(S_{11}), \ldots, \delta_n(S_{nm_n})\big)'$. The proposed vector ξ^p is accepted as the new state of the chain with probability

$$\alpha(\xi^c, \xi^p) = \min\left(1, \frac{p(\xi^p \mid \cdot) q(\xi^p, \xi^c)}{p(\xi^c \mid \cdot) q(\xi^c, \xi^p)}\right)$$

where $p(\xi \mid \cdot)$ is the full conditional for ξ (that is, the conditional distribution of ξ given all other parameters and the data).

For the parameters corresponding to the functions $g_0^{(h)}(t), \ldots, g_L^{(h)}(t)$ depending on time t, we adopt the computationally faster MH-algorithm based on conditional prior proposals. Unlike the algorithm based on IWLS proposals, this algorithm only requires evaluation of the log-likelihood, not of derivatives (see Fahrmeir and Lang (2001) for details). Note that the evaluation of derivatives would be particulary time-consuming for these parameters since further integrals are involved that have to be approximated numerically.

As the full conditionals of the variance parameters are (proper) inverse Gamma distributions, updating of hyperparameters can be done by simple Gibbs steps.

6 Application: Human sleep data

In this application, we analyse data on human sleep collected at the Max-Planck Institute for Psychiatry in Munich as a part of a larger study on sleep withdrawal. The part of the data we will consider is utilised to obtain a reference standard of the participants' sleeping behaviour at the beginning of the study. Therefore, the major goal is to obtain a valid description of the dynamics underlying the sleep process of the 70 participants of the study. For each of the patients, information on exactly one night is available.

Originally, the sleep process is recorded by electroencephalographic (EEG) measurements which are afterwards classified into the three states: awake, Non-REM and REM. The Non-REM state could be further differentiated but since our data set is comparably small, we will restrict ourselves to a three-state model. In addition to EEG measures taken every 30 seconds throughout the night, blood samples are taken from the patients approximately every 10 minutes, providing measurements on the nocturnal secretion of certain hormones, for example, cortisol. Including this covariate information in multi-state models allows to validate hypotheses about the relationship between the hormonal secretion level and changes in the transition intensities. For example, we will investigate whether an increased level of cortisol affects the transition intensities between Non-REM and REM-sleep phases, a relationship that has been found in exploratory correlation and variance analyses.

The general model structure we will consider is schematically represented in Figure 3. To obtain a somewhat simplified transition space, we aggregated the

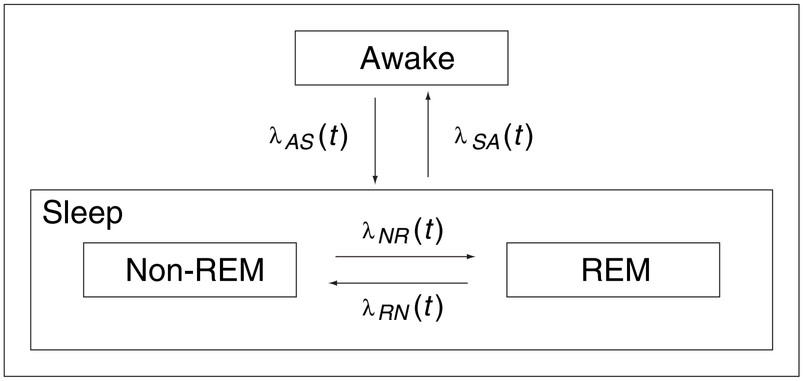

Figure 3 Schematic representation of sleep stages and the transitions of interest

transitions from awake to Non-REM and to REM, as well as the reverse transition into the single transitions awake to sleep and sleep to awake, respectively. Based on the previous considerations, we chose the following specification for the four remaining transition hazards:

$$\begin{aligned}
\lambda_{AS,i}(t) &= \exp\left[g_0^{(AS)}(t) + b_i^{(AS)}\right], \\
\lambda_{SA,i}(t) &= \exp\left[g_0^{(SA)}(t) + b_i^{(SA)}\right], \\
\lambda_{NR,i}(t) &= \exp\left[g_0^{(NR)}(t) + c_i(t)g_1^{(NR)}(t) + b_i^{(NR)}\right], \\
\lambda_{RN,i}(t) &= \exp\left[g_0^{(RN)}(t) + c_i(t)g_1^{(RN)}(t) + b_i^{(RN)}\right].
\end{aligned}$$

Each of the transitions is described in terms of a baseline effect $g_0^{(h)}(t)$ and a transition-specific frailty term $b_i^{(h)}$. In addition, we included time-varying effects $g_1^{(h)}(t)$ of high cortisol secretion for the transition rates between Non-REM and REM, where $c_i(t)$ is a dichotomised binary indicator for a high level of cortisol, that is, $c_i(t)$ takes the value one if the cortisol level exceeds 60 n mol/l at time t and zero otherwise. Therefore, the transition models between REM and Non-REM consist of two different intensity functions for a low level of cortisol ($g_0^{(h)}(t)$) and a high level of cortisol ($g_0^{(h)}(t) + g_1^{(h)}(t)$), respectively.

All time-varying effects $g_l^{(h)}(t)$, $l = 0, 1$ are modelled as cubic P-splines with second order difference penalty and 40 inner knots. We chose a relatively large number of knots to ensure enough flexibility of the time-varying functions. The transition- and patient-specific random effects $b_i^{(h)}$ are assumed to be i.i.d. Gaussian with $b_i^{(h)} \sim N(0, \tau_{hb}^2)$.

As a reference point we considered a purely parametric Markov model, where each of the transitions is assigned a time-constant rate not depending on any covariates. In this case, the maximum likelihood estimates of the transition intensities have a closed form and can be computed as the inverse of the average waiting time for a specific transition.

Estimated results for the time-varying baseline effects $g_0^{(h)}(t)$ together with (logarithmic) time-constant rates estimated from the parametric Markov model are displayed in Figure 4. Empirical Bayes inference and fully Bayesian inference lead to highly comparable results—with the transition from awake to sleep as a sole exception, where empirical Bayes inference yields a lower effect. Altogether we conclude that the transition rates are clearly varying over night with cyclic patterns for the transitions between awake and sleep, and the transition from Non-REM to REM. As was to be expected, the tendency to fall asleep again is particularly low for patients who wake up at the end of the night, that is, more than seven hours after sleep onset. In contrast, the tendency to wake up is roughly u-shaped and rather high in the beginning and especially high at the end of the night.

Concerning the transitions between Non-REM and REM sleep, the log-baseline effects $g_0^{(h)}(t)$ represent the effects for a low level of cortisol, while the $g_1^{(h)}(t)$ (compare Figure 5) describe deviations from these effects if the level of cortisol is high, that is, exceeds 60 n mol/l. In case of a low cortisol level, the intensity for a transition from Non-REM to REM is initially very low, but steeply increasing within the first hour after sleep onset followed by some ups and downs. In contrast, the intensity for the reverse transition from REM to Non-REM is highest immediately after sleep onset and afterwards decreases almost linearly. Figure 5 exhibits some additional time-variation for the transition rate from Non-REM to REM in case the level of cortisol is high. The additional effect of the reverse transition is less pronounced. Finally, frailty terms are identified for all transitions when applying fully Bayesian inference, while frailty terms are only identified for the transition from REM to Non-REM when applying empirical Bayes inference (results not shown).

The performance of the developed models is assessed using martingale residuals and compared to the parametric Markov process model. Figure 6 exemplarily displays martingale residuals for the transition from Non-REM to REM. Although the presentation as a time series plot is not very elucidating due to the accumulation of 70 individual processes, it allows to identify extreme outliers and to draw some general conclusions: The Markov model tends to overestimate the number of transitions (especially for the first hour after sleep onset), while the flexible, semiparametric models yield residuals with a relatively symmetric distribution about zero. In addition, the overall magnitude of the martingale residual processes is considerably smaller when inference is conducted fully Bayesian. This is due to the fact that the fully

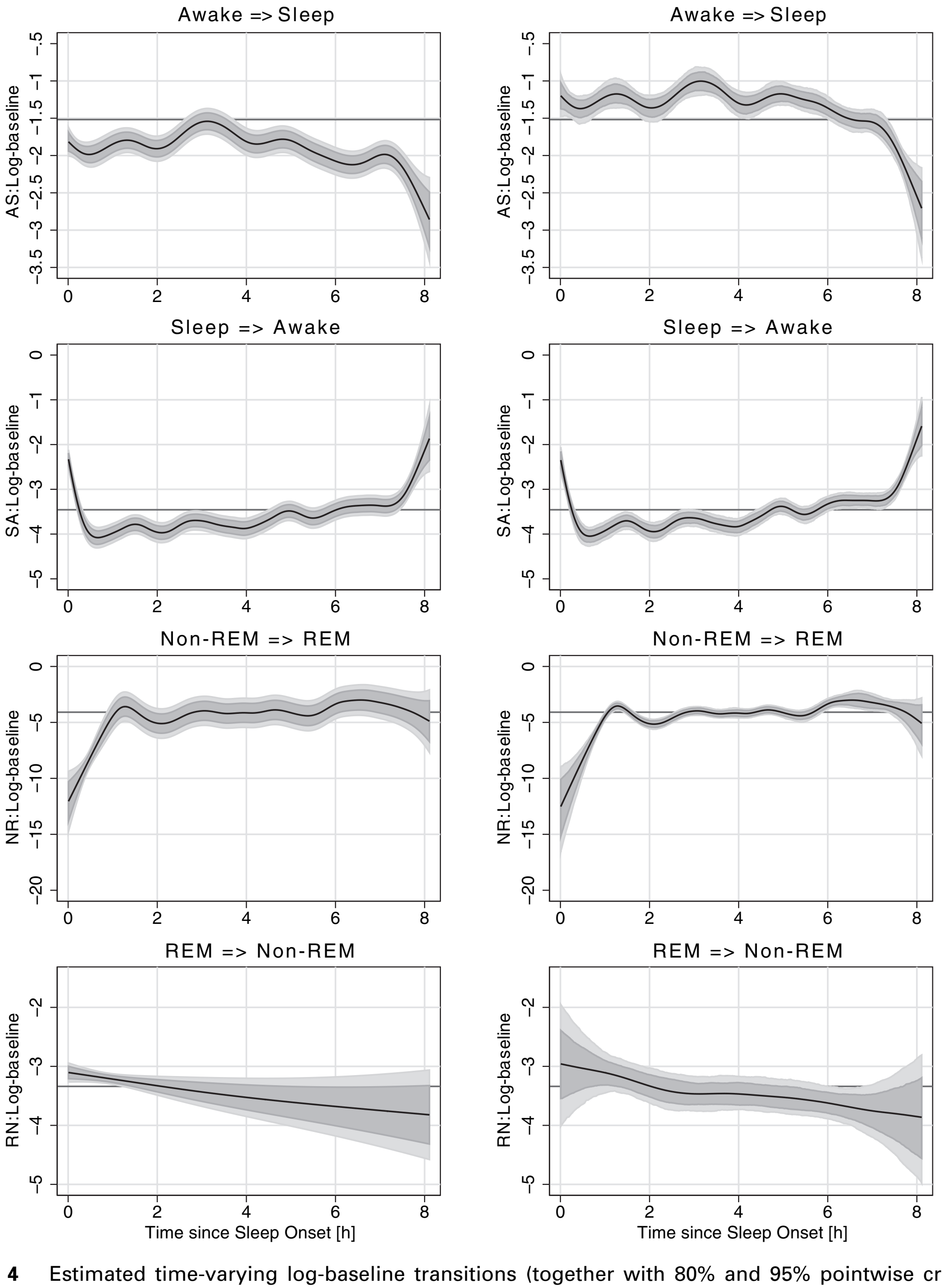

Figure 4 Estimated time-varying log-baseline transitions (together with 80% and 95% pointwise credible intervals) resulting from empirical Bayes (left panel) and fully Bayesian (right panel) inference. Horizontal grey lines mark time-constant estimates resulting from the Markov model

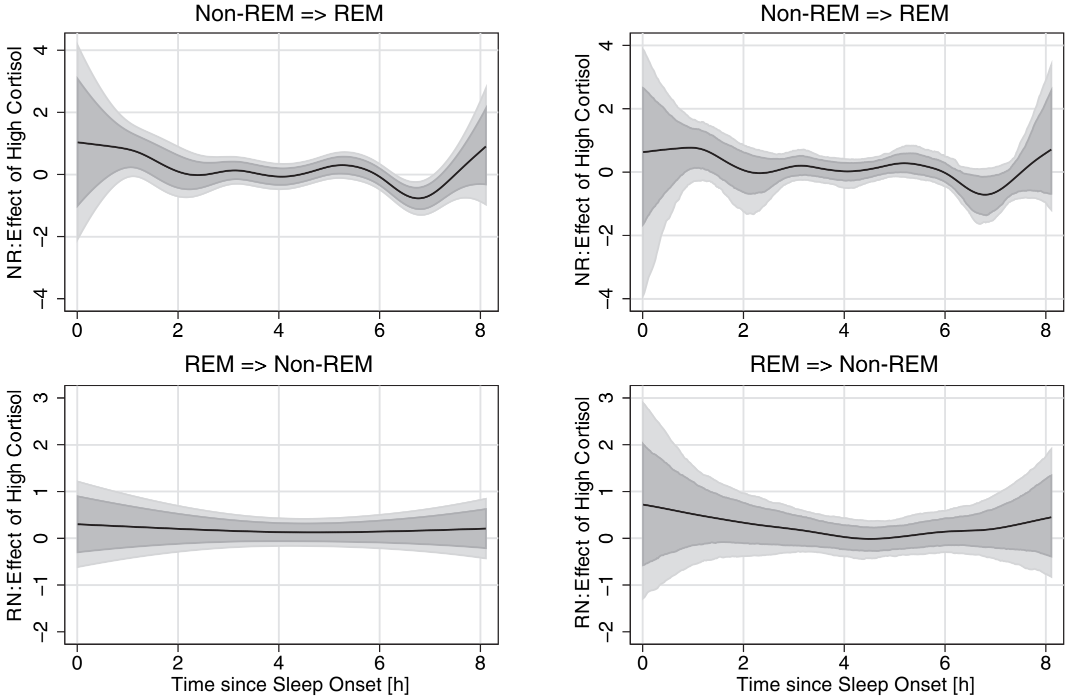

Figure 5 Estimated time-varying effects of high cortisol (together with 80% and 95% pointwise credible intervals) resulting from empirical Bayes (left panel) and fully Bayesian (right panel) inference

Bayesian frailty term estimates account for subject-specific differences which are ignored by the empirical Bayes estimates.

To gain additional insight into the distribution of the martingale residuals, Figure 7 displays kernel density estimates of the martingale residuals at selected time points. This illustration further supports the conclusion that semiparametric modelling of the transition intensities improves upon a purely parametric model. Overall, the fully Bayesian approach seems to perform best, with residual distributions which are mostly symmetric about zero. In contrast, the residual distributions for the parametric model are considerably shifted to either a positive or negative value, indicating under- and overestimation of the expected number of transitions, respectively. Results obtained with the empirical Bayes approach are somewhere in between fully Bayesian and parametric estimates. An exception is the transition from REM to Non-REM, where frailty terms are identified for both fully Bayesian and empirical Bayes inference, and hence both flexible models perform equally well. In summary, Figure 7 gives a further hint that individual-specific variation should be accounted for when modelling the transition intensities.

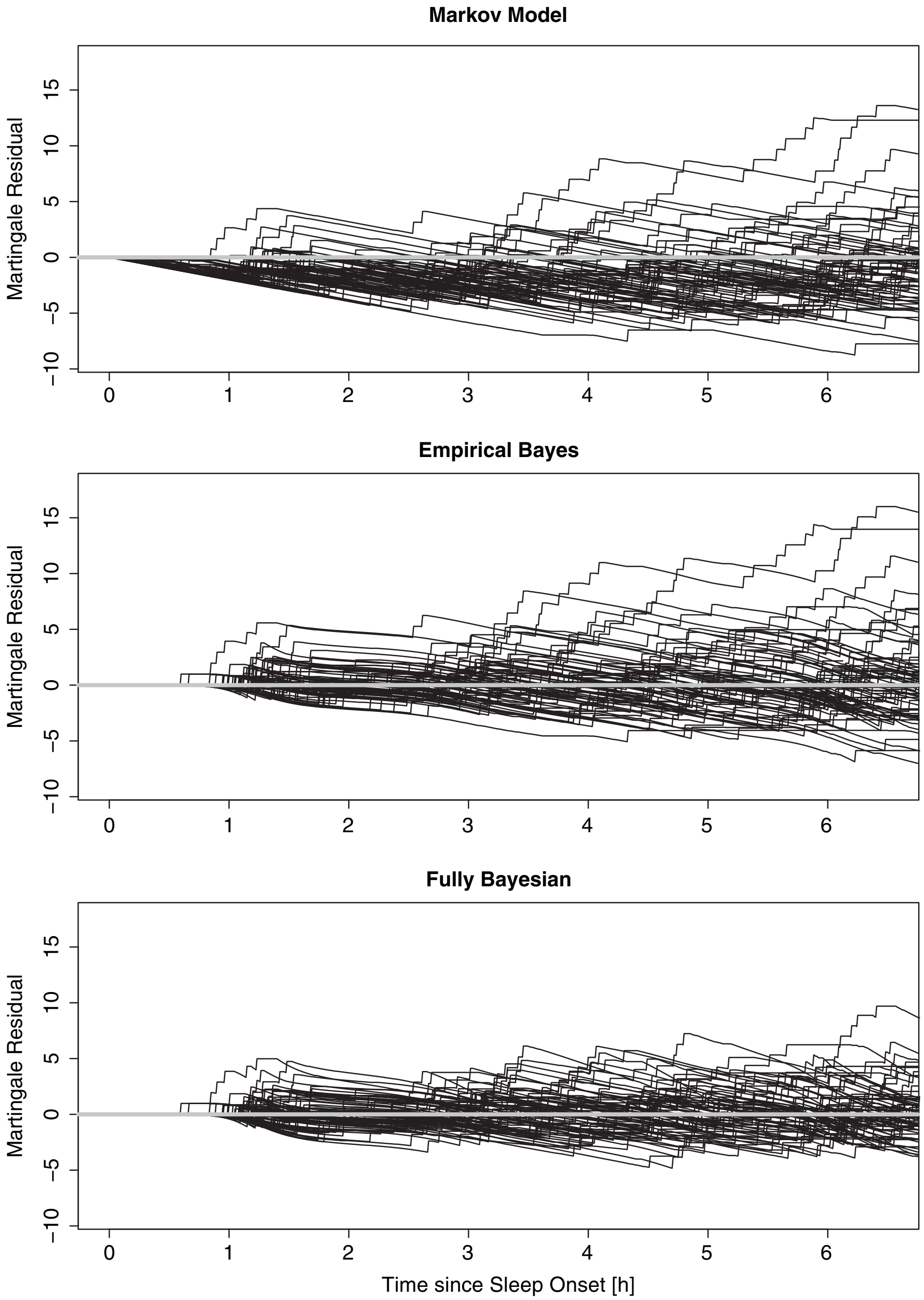

Figure 6 Martingale residuals for the transition from Non-REM to REM

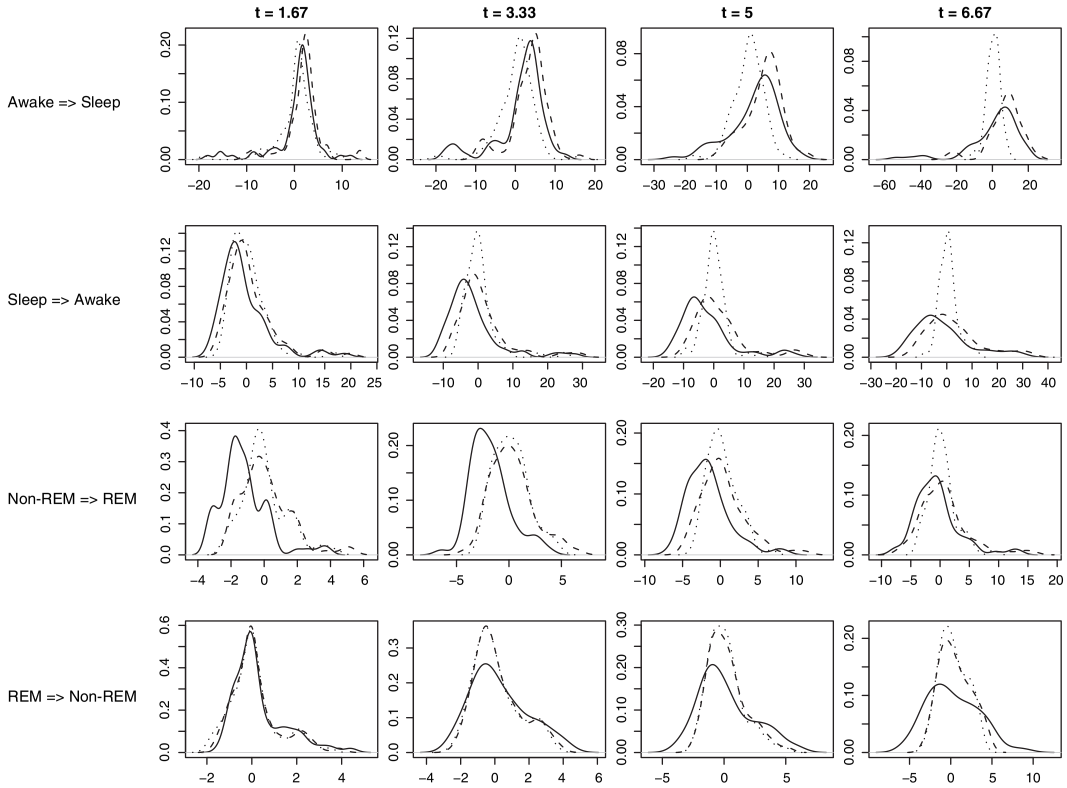

Figure 7 Kernel density estimates of martingale residuals at selected time points for the Markov model (solid lines), empirical Bayes inference (dashed lines) and fully Bayesian inference (dotted lines). x-axis: martingale residual, y-axis: density.

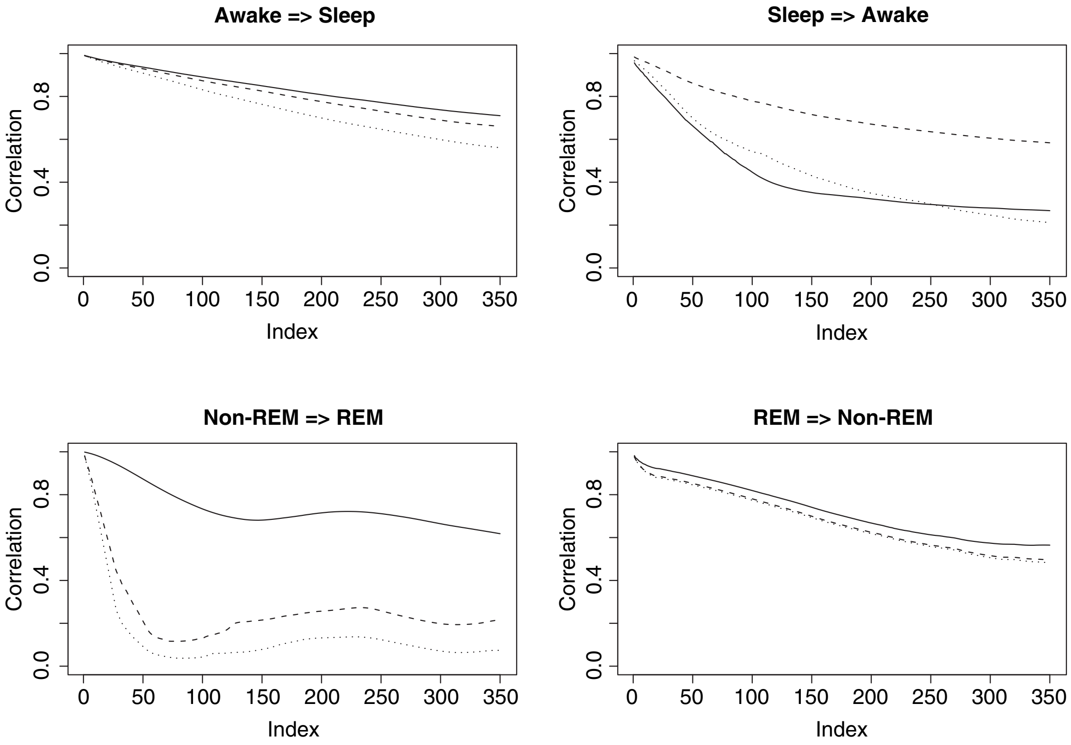

Figure 8 Autocorrelation functions for the Markov model (solid lines), empirical Bayes inference (dashed lines) and fully Bayesian inference (dotted lines)

Finally, Figure 8 shows empirical autocorrelation functions for 30 second increments of the residual processes. According to martingale theory, these increments should be (approximately) uncorrelated. Of course, it would be too strict to expect exactly uncorrelated residual processes but autocorrelations should die out quickly for a well-chosen model. Unfortunately, none of the models considered fully fulfills this requirement. In particular, the transitions from awake to sleep and from REM to Non-REM exhibit long-time autocorrelation and show only small differences between the three inferential procedures. In contrast, there is a clear improvement with the flexible model for the two remaining transitions. For the transition from Non-REM to REM, autocorrelations die out relatively quickly, especially for fully Bayesian estimates.

In summary, flexible models seem to improve upon the simple Markov model but are still not able to capture all of the essential features influencing the sleep process. However, since our data set only contains very little information about the participants of the sleep study, it probably is not very realistic to expect the model to fully explain the underlying dynamics. At least parts of the individual-specific variation are captured by the frailty effects which also proved to be important in the analysis of residual processes.

Of course, looking at the martingale residuals alone bears the risk to adjust the model too close to the observed data. Therefore, it would be useful to perform some out-of-sample prediction as discussed in Section 4. Since, however, our data set is quite small anyway, this strategy is not applicable in our case.

To investigate the sensitivity of MCMC-based estimates with respect to the prior assumptions, we recomputed our estimates with different hyperparameter settings for the inverse gamma priors of the variance parameters. More precisely, we considered the following combinations of hyperparameters: $a = b = 0.001$ (the default), $a = b = 0.0001$, $a = b = 0.00001$, $a = 1$ and $b = 0.001$, $a = -1$ and $b = 0$, $a = -0.5$ and $b = 0$. The latter two correspond to flat priors on the variance and the standard deviation, respectively. Note that in the limiting case, where $a = b = \epsilon \to 0$, the posterior is no longer proper. For small, positive values of ϵ, the posterior is still proper from a theoretical perspective but, in general, results tend to be too smooth. In contrast, relatively wiggly results are to be expected with the flat priors with $a < 0$.

Figure 9 shows some selected results from the sensitivity analysis: Log-baseline effects and time-varying effects for the default ($a = b = 0.001$) and the three most extreme hyperparameter settings ($a = b = 0.00001$, $a = 1$ and $b = 0.001$, $a = -1$, $b = 0$) are displayed on the same scale as estimates in Figures 4 and 5. From the results we can conclude the following: For the baseline effect for the transition between awake and sleep, there is some uncertainty about the heights of the maximal and minimal values of the effect. However, the differences are relatively small compared to the overall range of the effect and are also to be expected with priors enforcing a larger amount of smoothness. For the baseline of the transition intensity from Non-REM to REM, all results almost coincide except for the very end of the time-period. This simply reflects the fact that less transitions of this type are observed at this time. For the two further baseline effects and the random effects, results are qualitatively similar and no larger differences are observed. The most notable deviations between the prior specifications are to be found for the time-varying effects. For the transition from REM to Non-REM, there is uncertainty at the beginning of the night, due to the fact that the cortisol level is typically low at this time point. For the transition from Non-REM to REM, uncertainty in the estimates even leads to additional local minima and maxima for some special choices of hyperparameters allowing for more variation. Note that this sensitivity is adequately reflected in the credible intervals of the original estimates in Figure 5. For priors favouring smooth estimates, the time-varying effects are very close to straight lines.

In summary, although some sensitivity with respect to prior assumptions has been found for the time-varying effect, the results obtained with the standard choices seem to be quite reasonable. This conjecture is also supported by the results obtained with

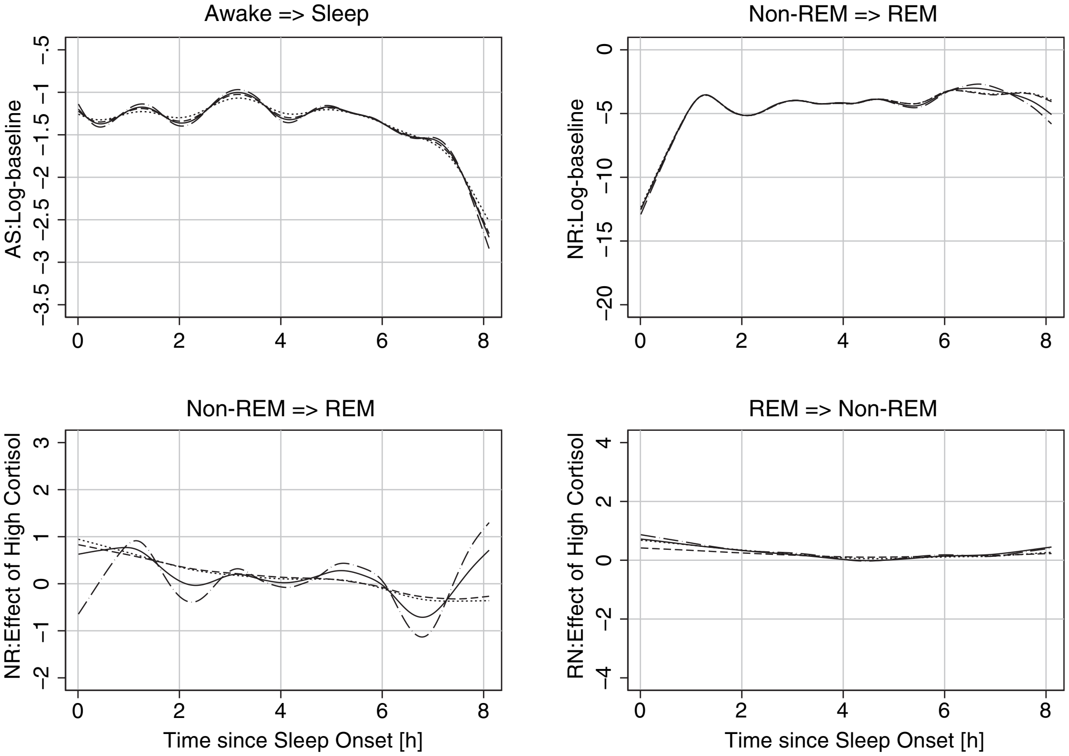

Figure 9 Selected results for the sensitivity analysis with hyperparameter settings $a = b = 0.001$ (—), $a = b = 0.00001$ (- - -), $a = 1$ and $b = 0.001$ (· · ·), $a = -1$, $b = 0$ (– · – ·)

the mixed model approach, which is inherently free from hyperparameter sensitivity (although it corresponds to one specific prior).

7 A small simulation study

Since the database in our application is quite small, we conducted a small simulation study to investigate the dependency of estimation results on the sample size. We considered a multi-state process that is build in analogy to the human sleep process, that is, the transition structure is given by the reachability graph in Figure 3.

Before discussing the set-up and the results of the simulation study in detail, we first present a simulation algorithm for general multi-state models. To simulate a multi-state model with a given set of hazard rates $\lambda_i^{(h)}(t)$, $h = 1, \ldots, H$, $i = 1, \ldots, n$, we require the following quantities:

(i) The cumulative distribution function of the $(r+1)$th transition time, conditional on the current state Y_{ir} and the rth transition time S_{ir}:

$$\begin{aligned} F_r(t) &= P(S_{i,r+1} \leq S_{ir} + t \mid Y_{ir}, S_{ir}) \\ &= P(T_{i,r+1} \leq t \mid Y_{ir}, S_{ir}) \\ &= 1 - \exp\left(-\int_{S_{ir}}^{S_{ir}+t} \alpha_i(t)dt\right) \\ &= 1 - \exp\left(-\left[A_i(S_{ir}+t) - A_i(S_{ir})\right]\right) \end{aligned} \tag{7.1}$$

where

$$\alpha_i(t) = \sum_{h=1}^{H} \alpha_i^{(h)}(t) = \sum_{h=1}^{H} I_i^{(h)}(t)\lambda_i^{(h)}(t), \quad S_{ir} < t \leq S_{i,r+1},$$

and

$$A_i(t) = \int_0^t \alpha_i(u)du.$$

Note that expression (7.1) extends the well-known formula for survival times that relates the cumulative distribution function $F(t)$ to the (cumulative) hazard rate $\Lambda(t)$ via $F(t) = 1 - \exp(-\Lambda(t))$. In fact, $\alpha_i(t)$ corresponds to the hazard rate in such a duration time model except for the fact that we have to account for the time point already reached by the process and the current state. The current state is only needed to ensure that the sum over the intensity processes involves only processes corresponding to transitions which are currently observable. Note that this is automatically accounted for in the aforementioned formula since the definition of the intensity processes includes the corresponding risk processes.

Inversion sampling allows to draw a random number from distribution (7.1). If $U \sim U[0,1]$, it follows that

$$F_r^{-1}(U) \sim F_r.$$

Expressing the cumulative distribution function in terms of $A_{ir}(t)$ and solving for t yields

$$T_{i,r+1} = A_{ir}^{-1}\left[-\log(1-U) + A_{ir}(S_{ir})\right] - S_{ir}.$$

Note that it is not required that the inverse of $A_{ir}(t)$ is available in closed form. For the simulation algorithm, numerical inversion will be sufficient. In fact, $A_{ir}(t)$ may also be approximated using numerical integration and inverted numerically in a second step.

(ii) The conditional probabilities

$$p_i^{(h)}(t) = P(\text{the transition from } Y_{ir} \text{ to } Y_{i,r+1} \text{ is of type } h \mid Y_{ir}, S_{i,r+1} = t) = \frac{\alpha_i^{(h)}(t)}{\sum_{h'=1}^{H} \alpha_i^{(h')}(t)}.$$

These probabilities are simply proportional to the values of the intensity processes $\alpha_i^{(h)}(t)$ at $S_{i,r+1} = t$, that is, the higher the intensity for a transition, the higher is the corresponding probability.

Based on these quantities, the simulation algorithm for the path of individual i up to a prespecified time $t_{\max}$ proceeds as follows:

(i) Generate an initial state Y_{i0}, either from external knowledge about the process or from an appropriate starting distribution.

(ii) Simulate the duration in the current state as

$$S_{i,r+1} = A_{ir}^{-1}(-\log(1-U) + A_{ir}(S_{ir})),$$

where $U \sim U[0,1]$. If $S_{i,r+1} > t_{\max}$, truncate $S_{i,r+1}$ to $t_{\max}$ and terminate the algorithm. Otherwise go to step iii.

(iii) Simulate the transition at time $S_{i,r+1}$ from the set of transitions based on the probabilities

$$p_i^{(h)}(t) = \frac{\alpha_i^{(h)}(S_{i,r+1})}{\sum_{h'=1}^{H} \alpha_i^{(h')}(S_{i,r+1})}, \quad h = 1, \ldots, H.$$

and go back to step (ii).

For the baseline hazard rates, we employed the following specifications:

$$\lambda_0^{(AS)}(t) = \begin{cases} 1.2\cos(t) + 2.5 & t \leq \pi \\ 1.3 & t > \pi \end{cases} \qquad \lambda_0^{(SA)}(t) = \begin{cases} 1.25\sin(t) + 2 & t \leq 2\pi \\ 2 & t > 2\pi \end{cases}$$

$$\lambda_0^{(NR)}(t) = 2.5\sin(1.5t) + 3.5 \qquad \lambda_0^{(RN)}(t) = 2\cos(t^2/4.5) + 2.5$$

To speed up and to simplify the simulation study, we did not consider time-varying effects. Based on the simulation algorithm, we generated 20 simulated data sets for different situations. First, we considered three different sample sizes, namely $n = 50$, $n = 100$ and $n = 200$. Based on the results from the 20 replications, we computed average point estimates. Figure 10 visualises the results for small and large sample sizes and the two estimation approaches (results for $n = 100$ are intermediate and therefore omitted). Note that we simulated the data on an hourly basis while the real

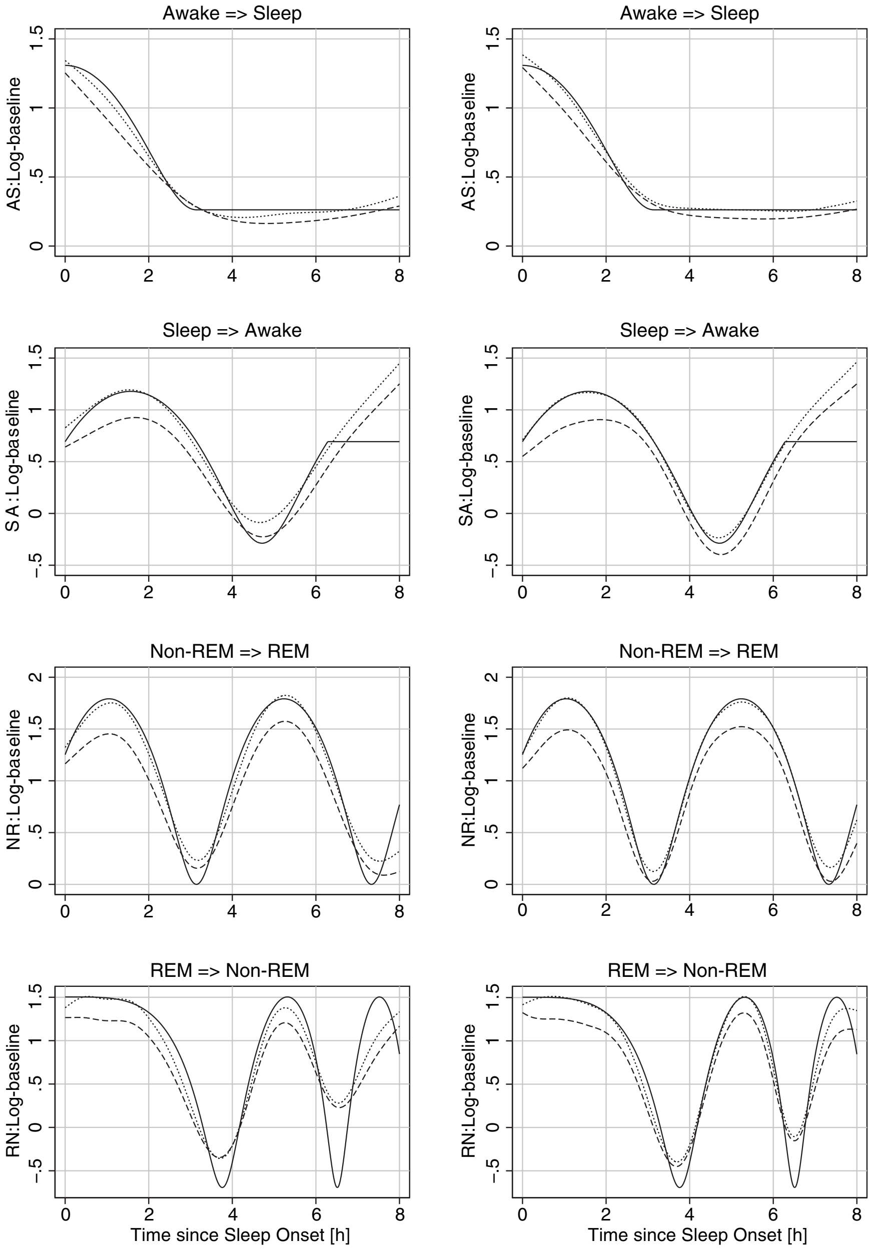

Figure 10 Simulation results for sample sizes $n = 50$ (left panel) and $n = 200$ (right panel). Mixed model based estimates are visualised as dashed line, MCMC based estimates as dotted line. The true values are included as solid line.

data in Section 6 were measured on a 30-seconds basis. Therefore, the magnitude of the baseline effects is different between the simulation data and the real data, but this does not affect the validity of our results.

Obviously, the differences between the sample sizes are not too large, although some improvement in the fit can be observed with $n = 200$. The differences between the mixed model based approach and the MCMC-based approach turn out to be larger: While both types of estimates reproduce the functional form of the hazard rate relatively well, the mixed model–based approach seems to have difficulties in capturing the overall level of the function. Note that this is consistent with the results for the transition from awake to sleep that we found in the application.

For all four transitions, the estimates reproduce quite well the functional form and the location of minima and maxima in the hazard rates. However, they also show the expected problems when abrupt changes in the curvature occur (transition from sleep to awake) or for highly fluctuating curves (transitions from REM to Non-REM and from Non-REM to REM). Still, the overall quality seems to be satisfactory when taking the complexity of the model and the small sample size into account.

In some further simulations we investigated the impact of random effects. The results confirmed our findings from the application: The mixed model–based approach typically is not able to detect individual-specific variability while the MCMC-based approach correctly identifies individual-specific effects. We also considered situations with a generally increased level of the hazard rates but this had only minor effect on the general quality of the estimation results.

8 Discussion

We have presented a computationally feasible semiparametric approach to the analysis of multi-state duration data motivated by an application to human sleep. Transition intensities are specified in a multiplicative manner in analogy to the Cox model, allowing for the inclusion of flexible nonparametric and time-varying effects. All parameters, including smoothing parameters, are estimated jointly using either an empirical Bayes or a fully Bayesian approach, therefore circumventing the need for subjective judgements. Some helpful tools for model validation and comparison have been considered on the basis of martingale residual processes.

When comparing the relative merits of the two proposed inferential procedures, MCMC-based estimation has the advantage of being structured modularly, thereby applying a divide-and-conquer strategy to the estimation problem. Dividing the full estimation problem in smaller parts allows to modify some of these parts without having the need to change the remaining ones, too. In large problems, this will also help to keep the computing time at a moderate level. However, in our small data set with a relatively simple structure of the transition intensities, the empirical Bayes approach was still much faster since it is not based on sampling techniques.

Thereby, it also avoids the necessity to validate mixing and convergence of a Markov chain, a task that becomes particularly cumbersome in complex models with a huge number of parameters. Since no priors have to be specified for the variance components in the empirical Bayes approach, the usual issue of sensitivity to the prior assumptions is also not present here. A drawback of mixed model–based empirical Bayes estimation is that parts of it rely on normal approximations. While these are usually not too problematic in the approximate marginal likelihood estimation of variance parameters, credible intervals for regression coefficients rely heavily on the assumption of asymptotic normality. In contrast, MCMC works with the posterior itself and therefore obtains more reliable interval estimates, provided that the Markov chain has converged. Therefore, MCMC seems to be better suited to small sample problems such as our sleep study. Here we also found that MCMC yielded somewhat preferable estimates, in particular with respect to individual-specific frailties where the database is even smaller than for the remaining regression effects.

The presented multi-state framework is easily extendable to different situations requiring more complicated modelling of covariate effects, such as spatial effects or interactions between covariates. In the future, application to such complicated data structures will be of particular interest to investigate the capabilities of Bayesian multi-state models. Of course, such extensions will require a larger database than in our application to make the effects well-identified.

A methodological extension will be the consideration of coarsened observations in analogy to interval censored survival data. This phenomenon is frequently observed in practice, in particular in medical applications where patients can be examined only at a prespecified, fixed set of time-points. In this case, the likelihood will in general not be available in analytic form, leading to additional numerical difficulties. In a fully Bayesian approach, the augmentation of true transition times in a data imputation step seems to be a promising alternative that avoids the computation of the exact likelihood.

Another direction of future research might be the consideration of different types of priors that, for example, allow for jumps or more abrupt changes in the log-baseline functions. Some of the estimates obtained in our application might suggest the necessity of such additional flexibility. Lang & Brezger (2004) present a modified prior, where an additional weight is introduced for the variance of the error terms of the random walk priors (2.3) and (2.4). This leads to varying amounts of smoothness over the co-domain of the modelled covariate. As an alternative, priors that mimic current regularisation penalties such as the LASSO may be considered.

Acknowledgments

The data analysed in this article have been kindly provided by Alexander Yassouridis, Max-Planck Institute of Psychiatry Munich. We would like to thank Ludwig Fahrmeir and Axel Munk for helpful discussions, as well as two anonymous referees whose comments were helpful in substantially improving upon a first version of this paper. Financial support by the German Science Foundation, Collaborative Research Center 386 'Statistical Analysis of Discrete Structures' is gratefully acknowledged.

References

Aalen OO, Fosen J, Weedon-Fekjær H, Borgan Ø and Husebye E (2004) Dynamic analysis of multivariate failure time data. *Biometrics*, **60**, 764–73.

Andersen PK, Borgan Ø, Gill RD and Keiding N (1993) *Statistical models based on counting processes*. New York: Springer.

Breslow NE and Clayton DG (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brezger A and Lang S (2006) Generalized additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, **50**, 967–91.

Crainiceanu C, Ruppert D, Claeskens G and Wand MP (2005) Exact likelihood ratio tests for penalised splines. *Biometrika*, **92**, 91–103.

Eilers PHC and Marx BD (1996) Flexible smoothing using B-splines and penalties (with comments and rejoinder). *Statistical Science*, **11**, 89–121.

Fahrmeir L and Klinger A (1998) A nonparametric multiplicative hazard model for event history analysis. *Biometrika*, **85**, 581–92.

Fahrmeir L, Kneib T and Lang S (2004) Penalized structured additive regression: a Bayesian perspective. *Statistica Sinica*, **14**, 731–61.

Fahrmeir L and Lang S (2001) Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society C*, **50**, 201–20.

Fahrmeir L and Kneib T (forthcoming) Propriety of posteriors in structured additive regression models: theory and empirical evidence. *Journal of Statistical Planning and Inference*.

Frühwirth-Schnatter S and Tüchler R (forthcoming) Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statistics and Computing*.

Gamerman D (1997) Efficient sampling from the posterior distribution in generalized linear models. *Statistics and Computing*, **7**, 57–68.

Greven S, Crainiceanu C, Küchenhoff H and Peters A (forthcoming) Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*.

Hall P and Heyde CC (1980) *Martingale limit theory and its application*. New York: Academic Press.

Hennerfeind A, Brezger A and Fahrmeir L (2006) Geoadditive survival models. *Journal of the American Statistical Association*, **101**, 1065–75.

Kneib T and Fahrmeir L (2007) A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, **34**, 207–28.

Lang S and Brezger A (2004) Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Ruppert D, Wand MP and Carroll RJ (2003) *Semiparametric regression*. Cambridge: Cambridge University Press.

Smith M and Kohn R (2002) Parsimoniuous covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, **94**, 777–94.

Yassouridis A, Steiger A, Klinger A and Fahrmeir L (1999) Modelling and exploring human sleep with event history analysis. *Journal of Sleep Research*, **8**, 25–36.