LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK
SONDERFORSCHUNGSBEREICH 386

Kneib, Hennerfeind:

Bayesian Semiparametric Multi-State Models

Sonderforschungsbereich 386, Paper 502 (2006)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Bayesian Semiparametric Multi-State Models

Thomas Kneib

Department of Statistics
Ludwig-Maximilians-University Munich
thomas.kneib@stat.uni-muenchen.de

Andrea Hennerfeind

Department of Statistics
Ludwig-Maximilians-University Munich
andreah@stat.uni-muenchen.de

Abstract

Multi-state models provide a unified framework for the description of the evolution of discrete phenomena in continuous time. One particular example are 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.

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 over time. Some common examples are depicted in Figure 1 in terms of their reachability graph for illustration. For recurrent events (Figure 1 (a)), the 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 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 'j', an individual can only move to contiguous states, i.e. either the disease gets worse and the individual moves to state '$j+1$', or the disease attenuates and the individual moves to state '$j-1$'. 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 about here.]

[Figure 2 about here.]

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 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 and it is well known that the dynamics of human sleep are strongly changing throughout the night with e.g. 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, Borgan, Gill, & Keiding (1993), and estimation is based on martingale theory. Fahrmeir & 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 & 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 (i.e. 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, Brezger & Fahrmeir (2006) and Kneib & Fahrmeir (2006) and allow for the simultaneous determination of all effects and smoothing parameters. Implementations will be available in release 1.5 of the software package BayesX, see `http://www.stat.uni-muenchen.de/~bayesx`.

As an illustration of our approach, we will 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. 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 (e.g. 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 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 and forms the basis for model validation based on martingale residuals. Section 4 presents details on the inferential schemes and Section 5 contains results of our application. The concluding Section 6 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 $\lambda_{hi}(t)$ where h, $h = 1, \ldots, k$, indexes the type of the transition and i, $i = 1, \ldots, n$, indexes the observations. 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, $\lambda_{hi}(t)$ is modelled in a multiplicative Cox-type way as

$$\lambda_{hi}(t) = \exp(\eta_{hi}(t)),$$

where

$$\eta_{hi}(t) = g_{h0}(t) + \sum_{l=1}^{L} g_{hl}(t)u_{il}(t) + \sum_{j=1}^{J} f_{hj}(x_{ij}(t)) + v_i(t)'\gamma_h + b_{hi} \tag{1}$$

is an additive predictor consisting of the following components:

- A time-varying, nonparametric baseline effect $g_{h0}(t)$ common for all observations.
- Covariates $u_{il}(t)$ with time-varying effects $g_{hl}(t)$. 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 $f_{hj}(x_{ij}(t))$ of continuous covariates $x_{ij}(t)$. For example, we might also include the hormonal level in a nonparametric way.
- Parametric effects γ_h of covariates $v_i(t)$.
- Frailty terms b_{hi} to account for unobserved heterogeneity.

After reindexing and suppressing the time-dependency of some of the quantities involved, we can represent the predictor vectors $\eta_h = (\eta_{h1}, \ldots, \eta_{hn})$ in generic notation as

$$\eta = V_1\xi_1 + \ldots + V_m\xi_m + V\gamma, \tag{2}$$

where V corresponds to the usual design matrix of fixed effects. The construction of the design matrices $V_1, \ldots, V_m$ 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 & 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, i.e.

$$f(x) = \sum_{j=1}^{d} \xi_j B_j(x).$$

Instead of estimating the resulting regression coefficients $\xi = (\xi_1, \ldots, \xi_d)'$ unrestricted, an additional penalty term is added to enforce smoothness of the estimated function. From a Bayesian perspective, this corresponds to a smoothness prior for ξ (Brezger & 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, i.e.

$$\xi_j = \xi_{j-1} + \varepsilon_j, \quad j = 2, \ldots, d,$$

for a first order random walk or

$$\xi_j = 2\xi_{j-1} - \xi_{j-2} + \varepsilon_j, \quad j = 3, \ldots, d,$$

for a second order random walk and Gaussian error terms $\varepsilon_j \sim N(0, \tau^2)$. 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.

In vector-matrix notation, penalised splines lead to the following representation for the function evaluations defining predictor (1): The baseline hazard rate can be expressed as $g_{h0}(t) = v'_{h0}\xi_{h0}$ where $v_{h0} = (B_{h01}(t), \ldots, B_{h0d}(t))'$ and $\xi_{h0} = (\xi_{h01}, \ldots, \xi_{h0d})'$. Similarly, we obtain $g_{hl}(t)u_l(t) = v'_{hl}\xi_{hl}$ for the time-varying effects with $v_{hl} =$

$(u_l(t)B_{hl1}(t), \ldots, u_l(t)B_{hld}(t))'$ and $f_{hj}(x_{ij}(t)) = v'_{hj}\xi_{hj}$ for nonparametric effects with $v_{hj} = (B_{hj1}(x_{ij}), \ldots, B_{hjd}(x_{ij}))'$. The design matrices in (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

$$p(\xi|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\xi' K \xi\right) \tag{3}$$

where the precision matrix $K = D'D$ is defined by the crossproduct of appropriate difference matrices D. Note, that in general distribution (3) is improper since K 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, i.e. $p(\gamma_h) \propto const$, and i.i.d. Gaussian priors $b_{hi} \sim N(0, \tau_h^2)$ with transition specific variances to the frailty terms. Note that these random effects distributions can also be cast into the multivariate form (3) by simply collecting all the random effects for one transition in the vector $\xi = (b_{h1}, \ldots, b_{hn})'$ and defining $K = I_n$. 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 4.1. 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 4.2.

3 Counting process representation and likelihood contributions

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

From classical counting process theory (see e.g. Andersen et al. (1993), Ch. VII.2), the intensity processes $\alpha_{hi}(t)$ of the counting processes $N_{hi}(t)$ are defined as the product of the hazard rate for type h transitions $\lambda_{hi}(t)$ and a predictable at-risk indicator process $Y_{hi}(t)$, i.e.

$$\alpha_{hi}(t) = Y_{hi}(t)\lambda_{hi}(t),$$

where the hazard rates are constructed in terms of covariates as described in Section 2. The at-risk indicator $Y_{hi}(t)$ takes the value one if individual i is at risk for a type h transition at time t and zero otherwise. For example, in the multi-state model of Figure 1a), 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

$$l_i = \sum_{h=1}^{k} \left[\int_0^{T_i} \log(\lambda_{hi}(t)) dN_{hi}(t) - \int_0^{T_i} \lambda_{hi}(t) Y_{hi}(t) dt \right], \tag{4}$$

where T_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 transition indicators $\delta_{hi}(t)$ and observed transition times t_{ij}, $j = 0, \ldots, n_i$, where $t_{i0} = 0$ and $t_{i,n_i} = T_i$. The indicators $\delta_{hi}(t)$ take the value one if a transition of type h is observed at time t for individual i and zero otherwise, while the t_{ij} are defined by the times at which the corresponding individual experiences a transition. This leads to the alternative log-likelihood formula

$$l_i = \sum_{j=1}^{n_i} \sum_{h=1}^{k} \left[\delta_{hi}(t_{ij}) \log(\lambda_{hi}(t_{ij})) - Y_{hi}(t_{ij}) \int_{t_{i,j-1}}^{t_{ij}} \lambda_{hi}(t) dt \right], \tag{5}$$

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 (4) reduces to a sum as shown in Equation (5) 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 utilised the trapezoidal rule due to its simplicity but of course more sophisticated methods could also be used if required.

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 Theorem to $N_{hi}(t)$ and obtain

$$\begin{aligned} N_{hi}(t) &= A_{hi}(t) + M_{hi}(t) \\ &= \int_0^t \alpha_{hi}(u) du + M_{hi}(t), \end{aligned}$$

where $M_{hi}(t)$ is a martingale and $A_{hi}(t)$ is a predictable process called the compensator of $N_{hi}(t)$. 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 $\hat{A}_{hi}(t)$ obtained from the model under consideration yields estimated residual processes $N_{hi}(t) - \hat{A}_{hi}(t)$. 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 & Heyde (1980), Sec. 1.6). Compare Section 5, where we will make use of martingale residuals in the context of our application.

4 Bayesian Inference

Based on the quantities considered in the last section 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 (3) will be treated as unknown constants which are to be estimated from their marginal likelihood. This will be facilitated by a mixed model representation of the predictor defining the transition hazards, see Section 4.1. 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 4.2. Both inferential procedures borrow from approaches which have been recently developed for continuous-time survival models (compare Kneib & Fahrmeir (2006) 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.

4.1 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 (3) can be represented as mixed models with i.i.d. random effects (compare e.g. Fahrmeir, Kneib & Lang (2004) or Ruppert, Wand & Carroll (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 the estimation of the variance parameters.

To be more specific, let ξ_j be the vector of regression coefficients describing a model term with $k_j = \text{rank}(K_j) \leq \dim(\xi_j) = d_j$. Our aim is to express ξ_j in terms of a k_j-dimensional vector of random effects b_j and a $(d_j - k_j)$-dimensional vector of fixed effects β_j. This can be achieved by applying the decomposition

$$\xi_j = \tilde{X}_j \beta_j + \tilde{Z}_j b_j \tag{6}$$

with suitably chosen design matrices $\tilde{X}_j$ and $\tilde{Z}_j$ of dimensions $(d_j \times d_j - k_j)$ and $(d_j \times k_j)$, respectively. The following conditions are assumed for the transformation in (6):

(i.) The compound matrix $(\tilde{X}_j \, \tilde{Z}_j)$ has full rank to make (6) a one-to-one transformation.

(ii.) $\tilde{X}_j' K_j = 0$ yielding a flat prior for β_j, i.e. β_j can be interpreted as a vector of fixed effects.

(iii.) $\tilde{Z}_j' K_j \tilde{Z}_j = I_{k_j}$ yielding an i.i.d. Gaussian prior for b_j, i.e. $b_j \sim N(0, \tau^2 I_{k_j})$ can be interpreted as a vector of i.i.d. random effects with variance τ^2.

Correspondingly the vector of function evaluations transforms to

$$V_j \xi_j = V_j(\tilde{X}_j \beta_j + \tilde{Z}_j b_j) = X_j \beta_j + Z_j b_j$$

with $X_j = V_j \tilde{X}_j$ and $Z_j = V_j \tilde{Z}_j$. 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, compare the discussion in Kneib & Fahrmeir (2006). Each of the nonparametric effects in a transition intensity yields a column of ones in the design matrix X_j which models the overall effects of the corresponding function. To obtain a valid model specification, we include an intercept in

each of the transition intensities 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 $ug(t)$.

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 instead refer to the complete description in Kneib & Fahrmeir (2006).

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

$$l_p(\beta, b, \tau^2) = \sum_{i=1}^{n} l_i + \sum_j \frac{1}{2\tau_j^2} b_j' b_j, \tag{7}$$

where β, b and τ^2 are vectors collecting all fixed effects, random effects and variances, respectively. The first term in (7) corresponds to the likelihood (4) while the second term corresponds to a sum over all prior distributions in the model. Since (7) 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.

The variances themselves are to be obtained from the marginal posterior, i.e. by maximising (7) after integrating out all regression coefficients:

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

Of course this integral can hardly be solved analytically or numerically in practice since β and b will be high-dimensional. Therefore we apply a Laplace approximation to $l_{marg}(\tau^2)$ similar as in Breslow & 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 likelihood allows to device 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.

4.2 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. Hence, hyperparameters are an integral part of the model and will be estimated jointly with all other parameters using MCMC techniques. We routinely assign inverse Gamma priors $IG(a_j; b_j)$

$$p(\tau_j^2) \propto \frac{1}{(\tau_j^2)^{a_j+1}} \exp\left(-\frac{b_j}{\tau_j^2}\right) \tag{8}$$

to all variances. They are proper for $a_j > 0$, $b_j > 0$, and we use $a_j = b_j = 0.001$ as a standard choice for a weakly informative prior. Note that uniform priors are a special (improper) case of the prior (8) with $a_j = -0.5$, $b_j = 0$, still leading to proper posteriors under regularity assumptions. 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 and τ^2 the vector of all variance parameters. Full 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 factorized 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 λ_{hi}), given the rest of 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_j, as well as fixed effects γ and frailty terms b, 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 & Lang (2006). Suppose we want to update a certain parameter vector ξ_j, with current value ξ_j^c of the chain. Then a new value ξ_j^p is proposed by drawing a random vector from a (high-dimensional) multivariate Gaussian proposal distribution $q(\xi_j^c, \xi_j^p)$, which is obtained from a quadratic approximation of the log-likelihood by a second order Taylor expansion with respect to ξ_j^c, in analogy to IWLS iterations in generalized 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

$$P_j = V_j' W(\xi_j^c) V_j + \frac{1}{\tau_j^2} K_j, \quad m_j = P_j^{-1} V_j' W(\xi_j^c)(\tilde{y} - \eta_{-j}).$$

Here, $\eta_{-j} = \eta - V_j \xi_j$ is the part of the linear predictor associated with all remaining effects in the model and $W(\xi_j^c) = \text{diag}(w_{11}, \ldots, w_{1n_1}, \ldots, w_{nn_n})$ is the weight matrix for IWLS with weights calculated from the current state ξ_j^c as $w_{ik} = \int_{t_{i,k-1}}^{t_{ik}} Y_i(u)\lambda_i(u)du$ for $i = 1, \ldots, n$, $k = 1, \ldots, n_i$. The vector of working observations $\tilde{y}$ is given by

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

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

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

where $p(\xi_j \mid \cdot)$ is the full conditional for ξ_j (i.e. the conditional distribution of ξ_j given all other parameters and the data).

For the parameters corresponding to the functions $g_0(t), \ldots, g_L(t)$ 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 & Lang (2001) for details). Note that the evaluation of derivatives would be particulary time-consuming for these parameters since further integrals are involved that are not to be solved by explicit formulae.

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

5 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.

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, e.g. cortisol. Including this covariate information in multi-state models allows to validate assumptions 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 effects the transition intensities between Non-REM and REM-sleep phases, a relationship that has been found in exploratory correlation and variance analyses.

[Figure 3 about here.]

The general model structure we will consider is schematically represented in Figure 3. To obtain a somewhat simplified transition space, we aggregated the 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 dichotomized binary indicator for a high level of cortisol, i.e. $c_i(t)$ takes the value one if the cortisol level exceeds 60 n mol/l at time t and zero otherwise. Therefore, each transition model consists of two different intensity functions for a low level of cortisol ($g_0(t)$) and a high level of cortisol ($g_0(t) + g_1(t)$) respectively.

All time-varying effects $g_j^{(h)}$, $j = 0, 1$ 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 $g_0^{(h)}$ 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 awake at the end of the night, i.e. after more than 7 hours after sleep onset. By 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 $g_0^{(h)}$, $h = NR, RN$ mark the effects for a low level of cortisol, while $g_1^{(h)}$ (compare Figure 5) describe deviations from these effects if the level of cortisol is high, i.e. 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).

[Figure 4 about here.]

[Figure 5 about here.]

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 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 via both, fully Bayesian and empirical Bayes inference, and hence both flexible models perform equally good. Hence, Figure 7 gives a further hint that individual-specific variation should be accounted for when modelling the transition intensities.

[Figure 6 about here.]

[Figure 7 about here.]

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.

[Figure 8 about here.]

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 is captured by the frailty effects which also proved to be important in the analysis of residual processes.

6 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 were 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 were 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.

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 more 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 data base 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.

Acknowledgments: The data analysed in this article have been kindly provided by Alexander Yassouridis, Max-Planck-Institute of Psychiatry Munich. Thanks are given to Ludwig Fahrmeir and Axel Munk for helpful discussions. Financial support by the German Science Foundation, Collaborative Research Center 386 "Statistical Analysis of Discrete Structures" is gratefully acknowledged.

References

Aalen, O. O., Fosen, J., Weedon-Fekjær, H., Borgan, Ø& Husebye, E. (2004). Dynamic analysis of multivariate failure time data. *Biometrics*, **60**, 764-773.

Andersen, P. K., Borgan, Ø, Gill, R. D. & Keiding, N. (1993) *Statistical Models Based on Counting Processes,* Springer.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

Brezger, A. & Lang, S. (2006). Generalized additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, **50**, 967-991.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing using B-splines and penalties (with comments and rejoinder). *Statistical Science* **11**, 89-121.

Fahrmeir, L. & Klinger, A. (1998). A nonparametric multiplicative hazard model for event history analysis. *Biometrika*, **85**, 581-592.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective. *Statistica Sinica*, **14**, 731-761.

Fahrmeir, L. & Lang, S. (2001). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Journal of the Royal Statistical Society C*, **50**, 201-220.

Gamerman, D. (1997). Efficient Sampling from the Posterior Distribution in Generalized Linear Models. *Statistics and Computing*, **7**, 57-68.

Hall, P. & Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*, Academic Press, New York.

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, **101**, 1065-1075.

Kalbfleisch, J. D. & Prentice, R. L. (2002) *The Statistical Analysis of Failure Time Data,* Wiley.

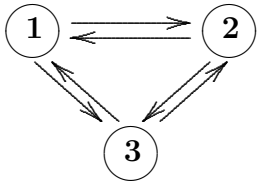
Kneib, T. & Fahrmeir, L. (2006) A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, to appear.

Kneib, T. (2006) Geoadditive hazard regression for interval censored survival times. *Computational Statistics and Data Analysis*, **51**, 777-792.

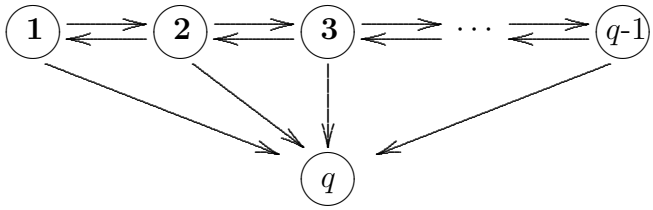
Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression*, Cambridge University Press.

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

(a) Recurrent events



(b) Disease progression



(c) Competing risks

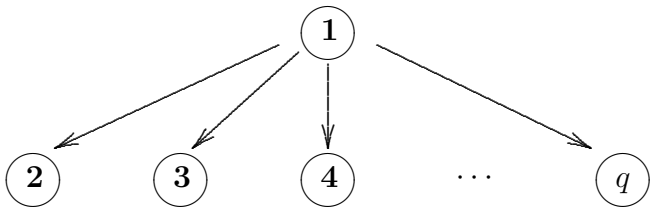


Figure 1: Reachability graphs of some common multi-state models.

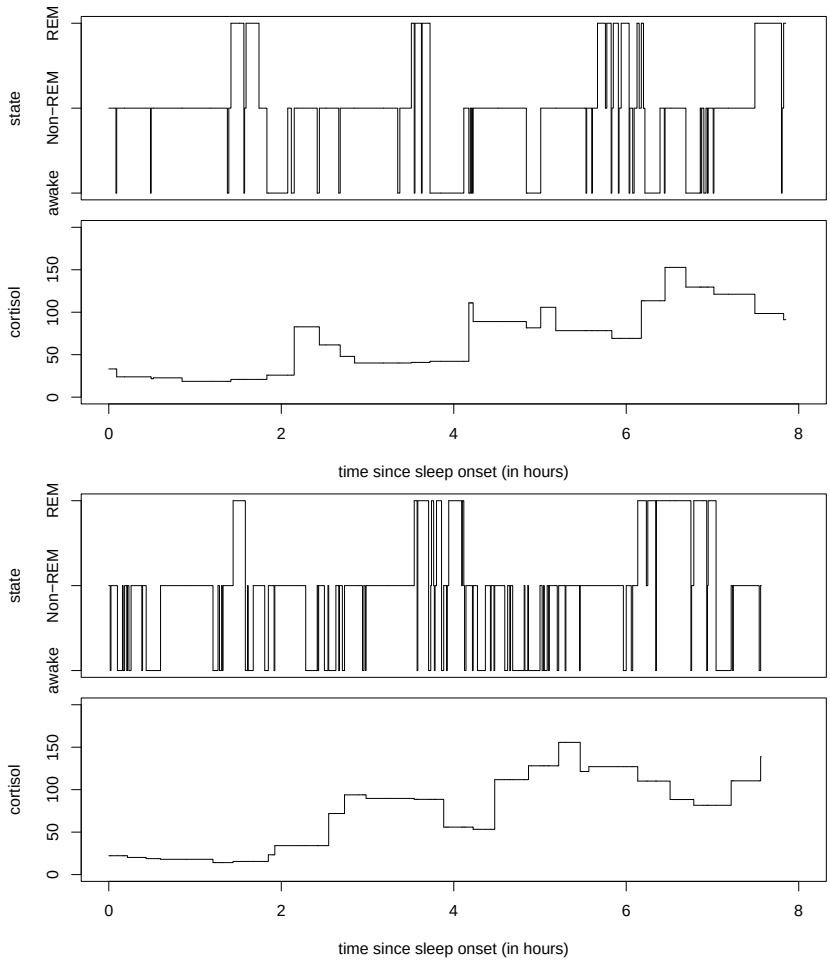


Figure 2: Realisations of two individual sleep processes and corresponding nocturnal cortisol secretion.

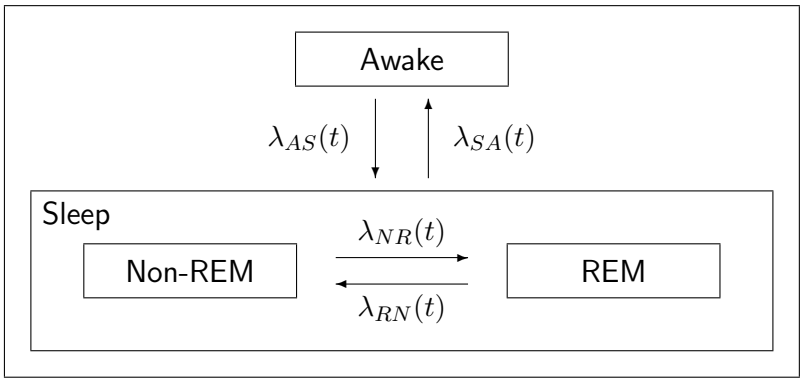


Figure 3: Schematic representation of sleep stages and the transitions of interest.

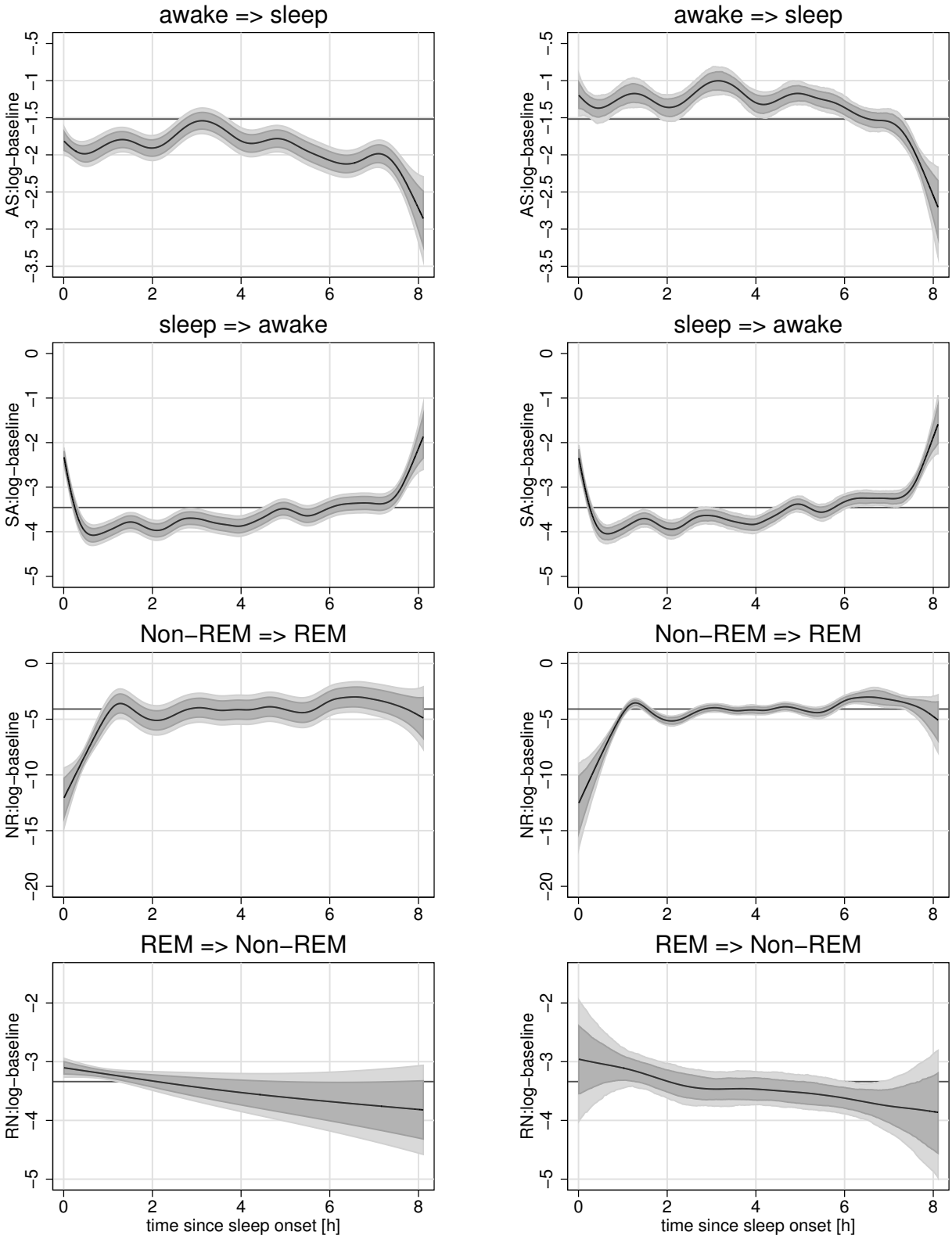


Figure 4: Estimated time-varying log-baseline transitions (together with 80% and 95% pointwise credible intervals) resulting from empirical Bayes (left panel) and fully Bayesian (right panel) inference. Horizontal grey lines mark time-constant estimates resulting from the Markov model.

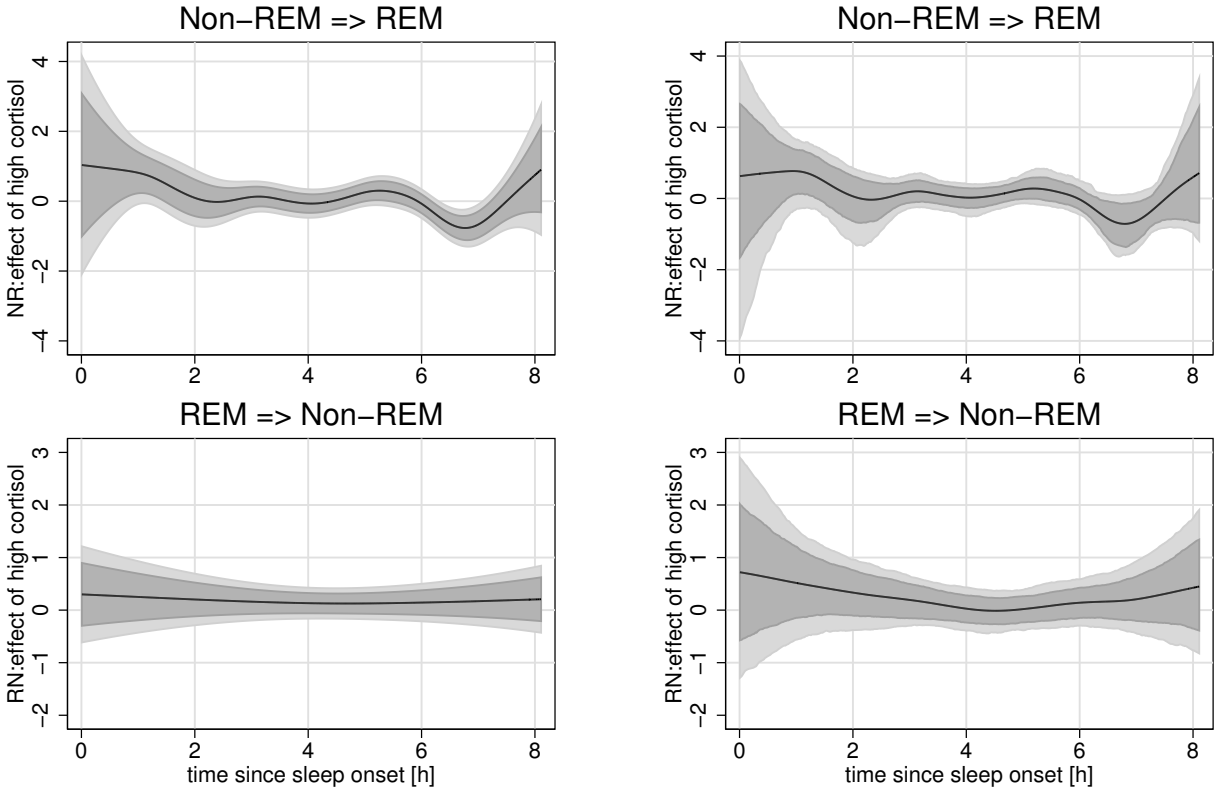


Figure 5: Estimated time-varying effects of high cortisol (together with 80% and 95% pointwise credible intervals) resulting from empirical Bayes (left panel) and fully Bayesian (right panel) inference.

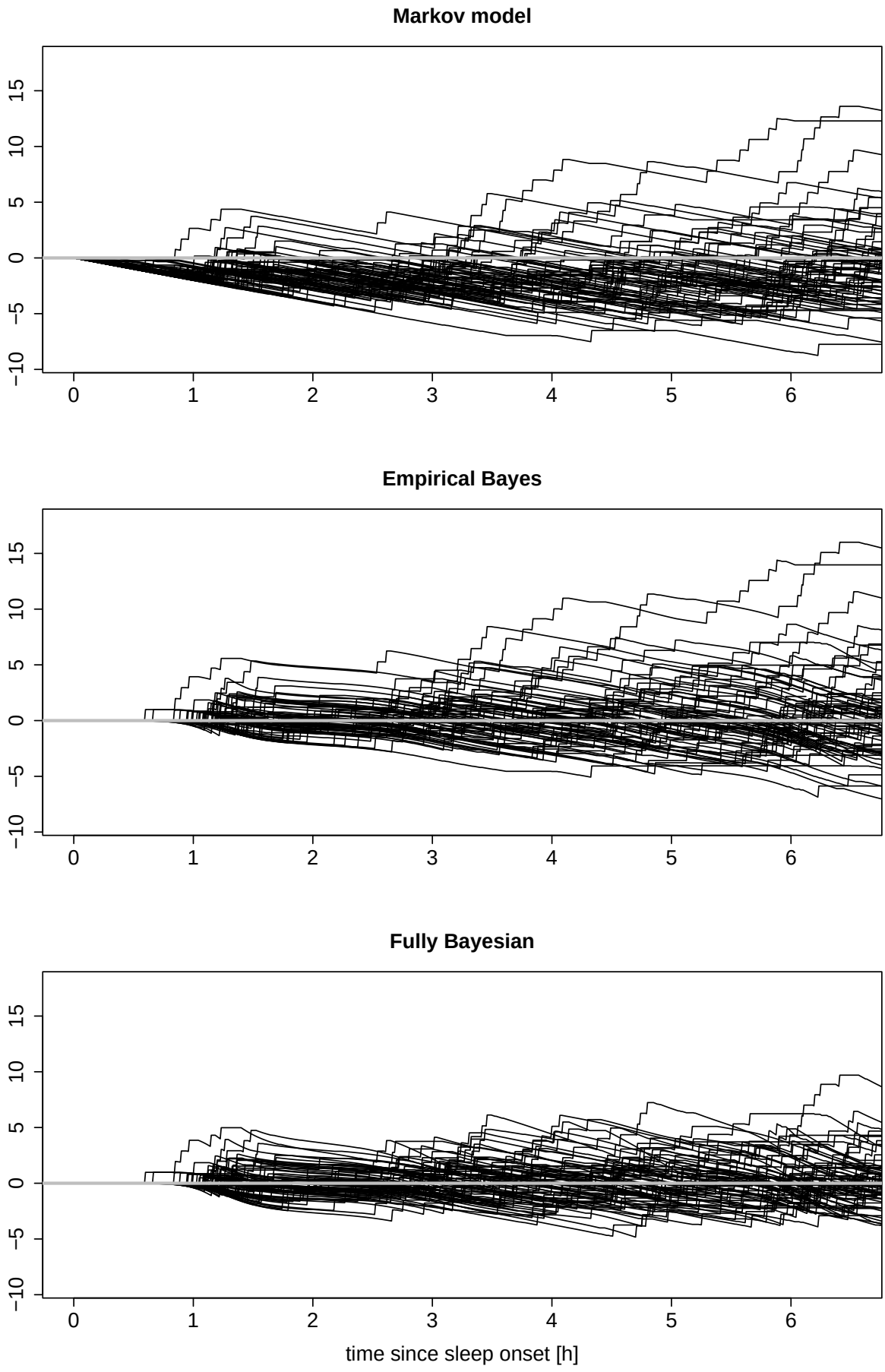


Figure 6: Martingale residuals for the transition from Non-REM to REM.

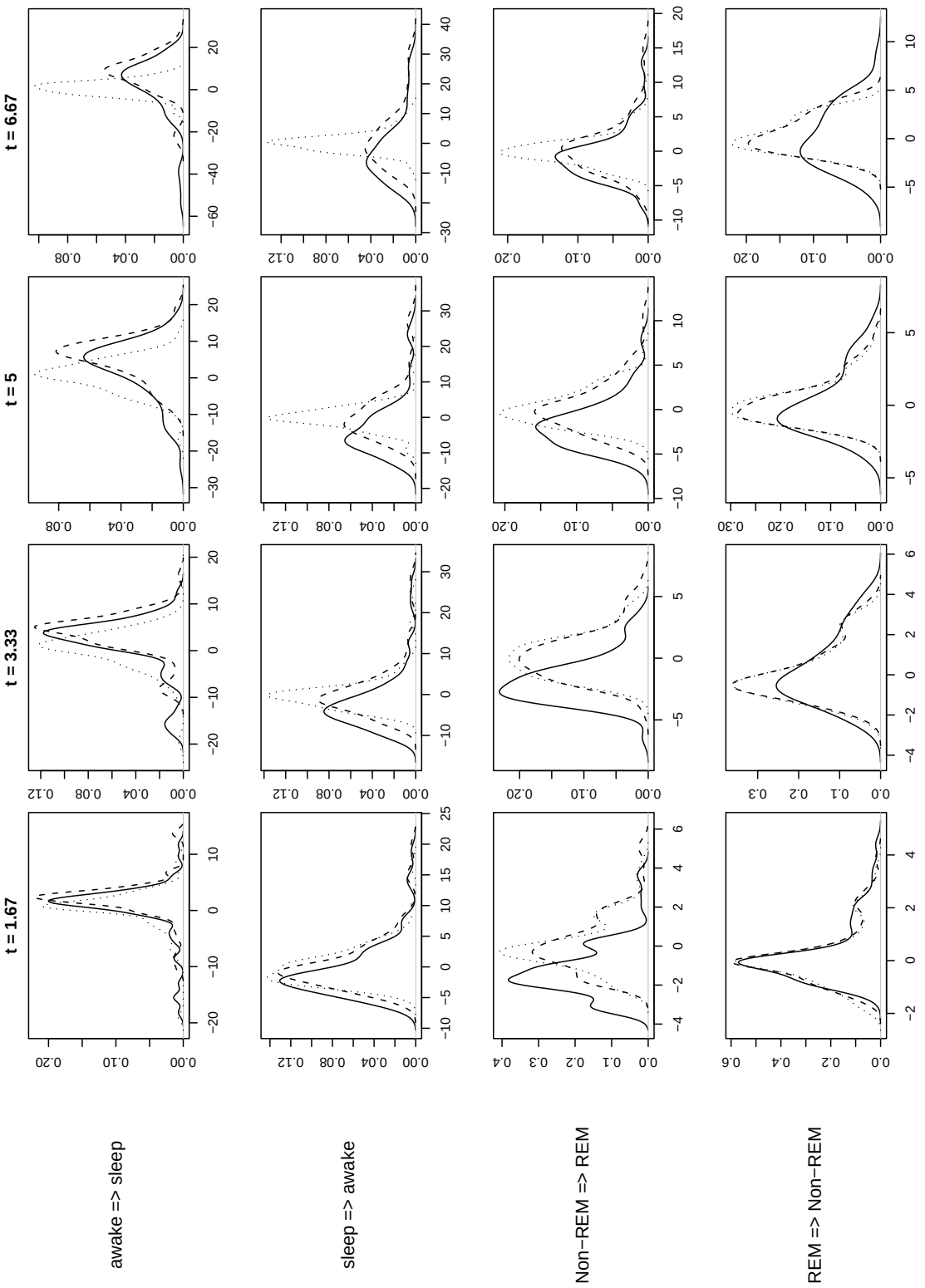


Figure 7: Kernel density estimates of martingale residuals at selected points of time for the Markov model (solid lines), empirical Bayes inference (dashed lines) and fully Bayesian inference (dotted lines).

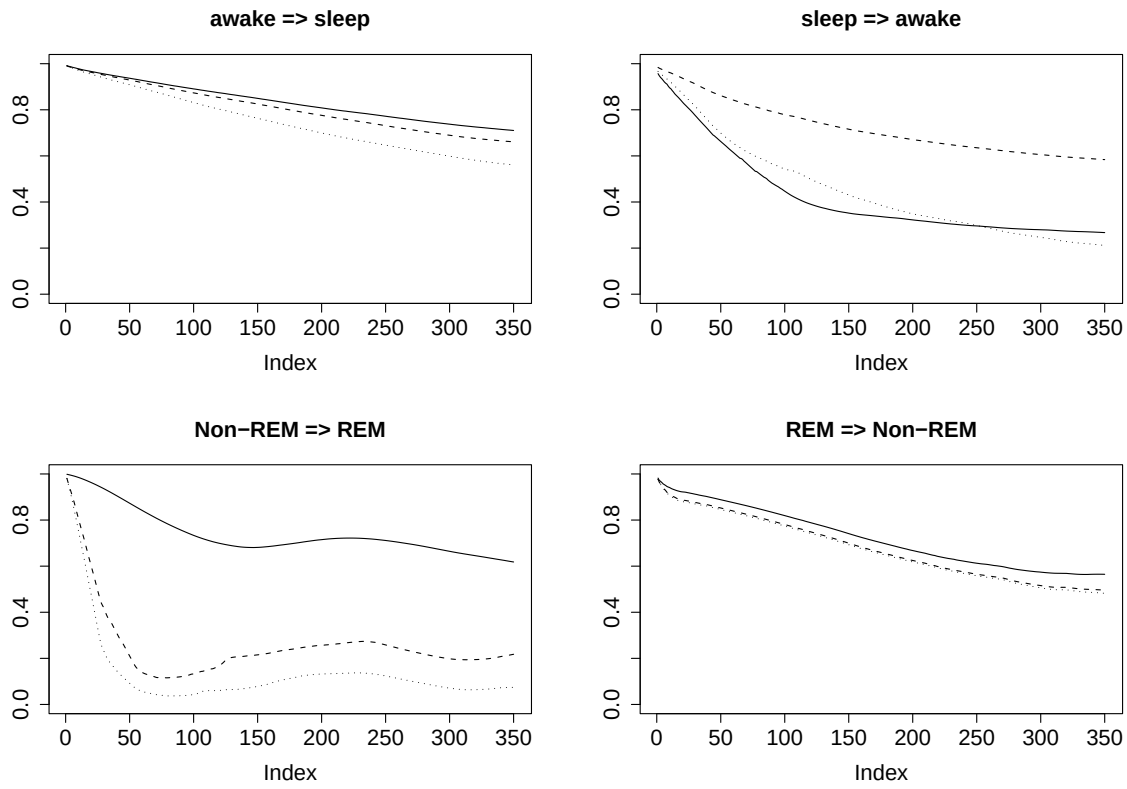


Figure 8: Autocorrelation functions for the Markov model (solid lines), empirical Bayes inference (dashed lines) and fully Bayesian inference (dotted lines).