Empirical Bayes Inference in Structured Hazard Regression

Thomas Kneib

Department of Statistics
Ludwig-Maximilians-University Munich

20.11.2006

Outline

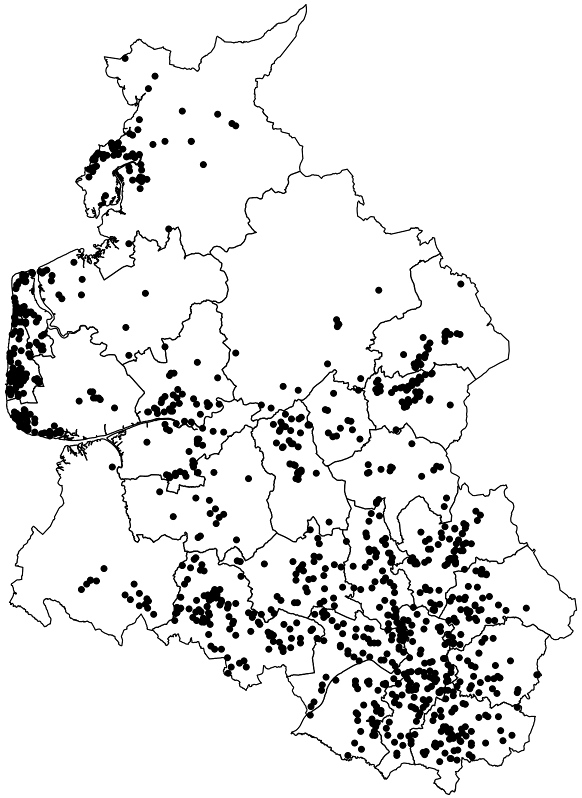
- Leukemia survival data.
- Structured hazard regression for continuous survival times.
- Empirical Bayes inference in structured hazard regression.
- Multi-state models.

Leukemia Survival Data

- Survival times of adults after diagnosis of acute myeloid leukemia.
- 1,043 cases diagnosed between 1982 and 1998 in Northwest England.
- 16 % (right) censored.
- Continuous and categorical covariates:

age	age at diagnosis,
wbc	white blood cell count at diagnosis,
sex	sex of the patient,
tpi	Townsend deprivation index.

- Spatial information in different resolution.



- Classical Cox proportional hazards model:

$$\lambda(t;x) = \lambda_0(t)\exp(x'\gamma).$$

- Baseline hazard $\lambda_0(t)$ is a nuisance parameter and remains unspecified.
- Estimate γ based on the partial likelihood.
- Questions / Limitations:
  - Simultaneous estimation of baseline hazard rate and covariate effects.
  - Flexible modelling of covariate effects (e.g. nonlinear effects, interactions).
  - Spatially correlated survival times.
  - Non-proportional hazards models / time-varying effects.

$\Rightarrow$ Structured hazard regression models.

- Replace usual parametric predictor with a flexible semiparametric predictor

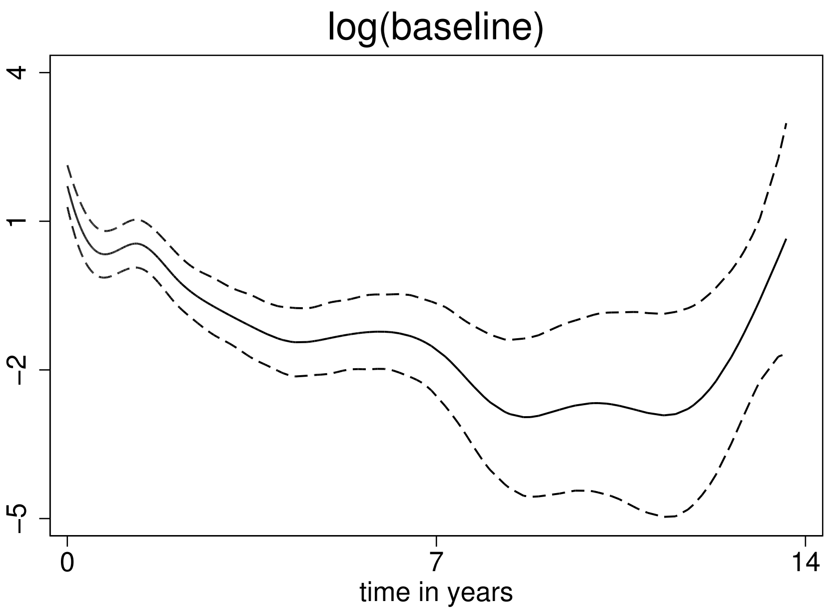
$$\lambda(t;\cdot) = \lambda_0(t)\exp[g_1(t)sex + f_1(age) + f_2(wbc) + f_3(tpi) + f_{spat}(s_i)]$$

and absorb the baseline

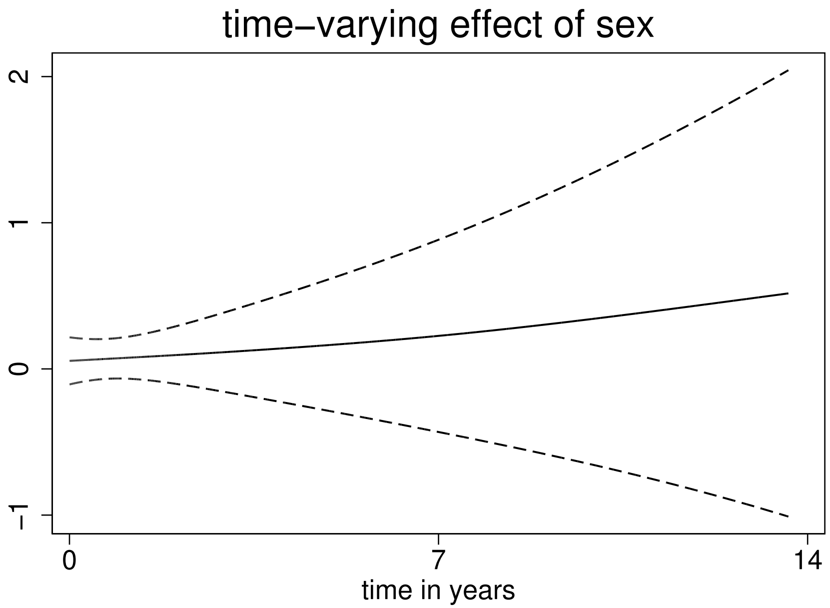
$$\lambda(t;\cdot) = \exp[g_0(t) + g_1(t)sex + f_1(age) + f_2(wbc) + f_3(tpi) + f_{spat}(s_i)]$$

where

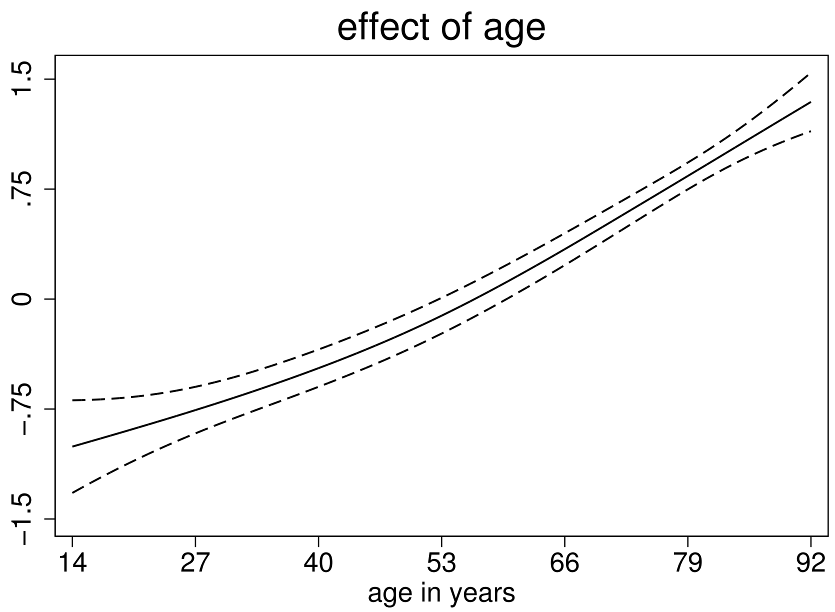
- $g_0(t) = \log(\lambda_0(t))$ is the log-baseline hazard,
- $g_1(t)$ is a time-varing gender effect,
- f_1, f_2, f_3 are nonparametric functions of age, white blood cell count and deprivation, and
- f_{spat} is a spatial function.



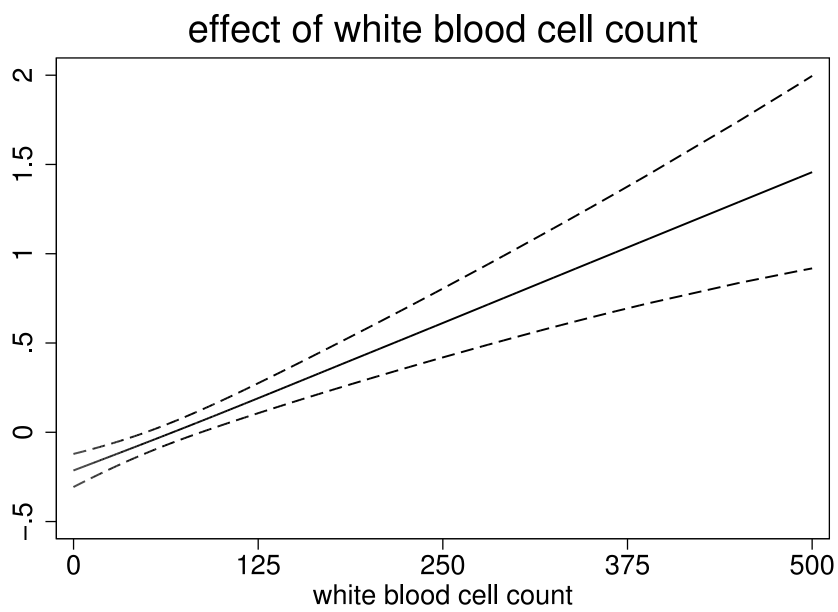
Log-baseline hazard.



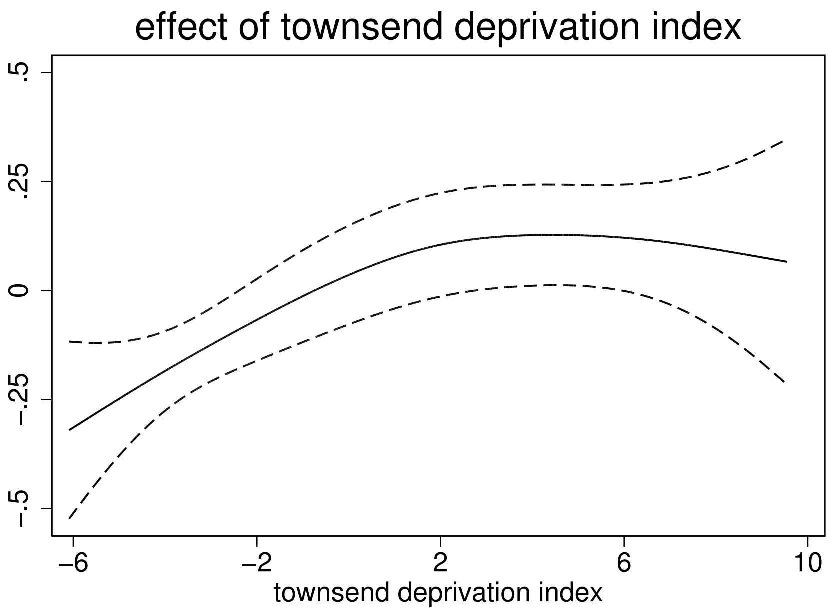
Time-varying gender effect.



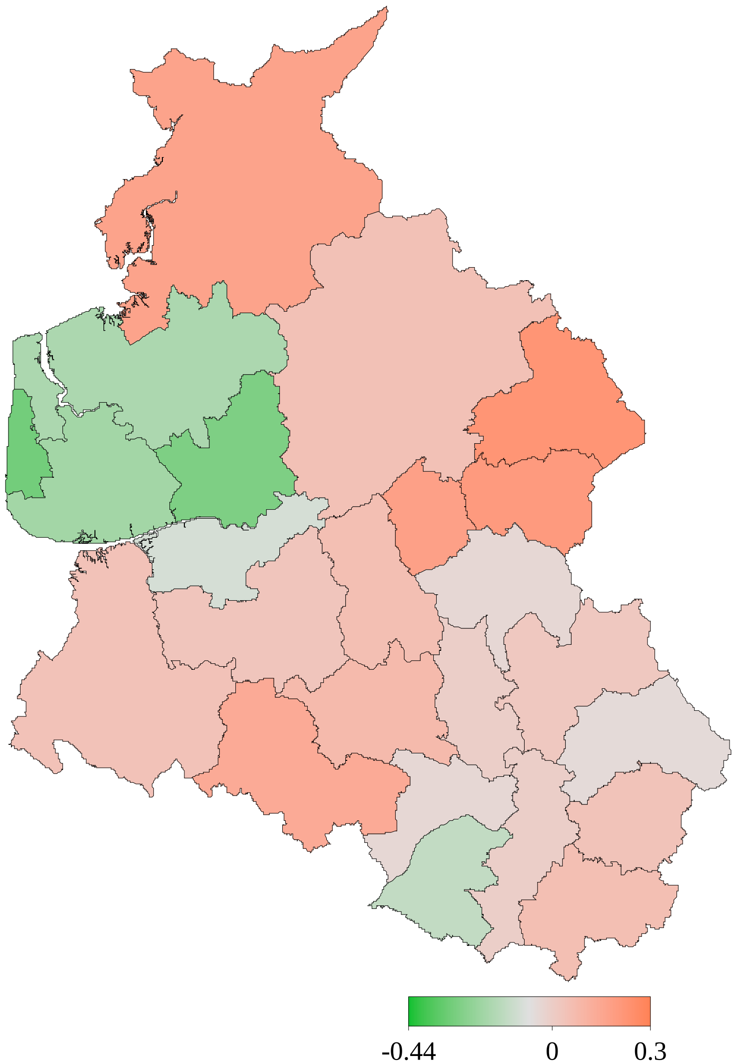
Effect of age at diagnosis.



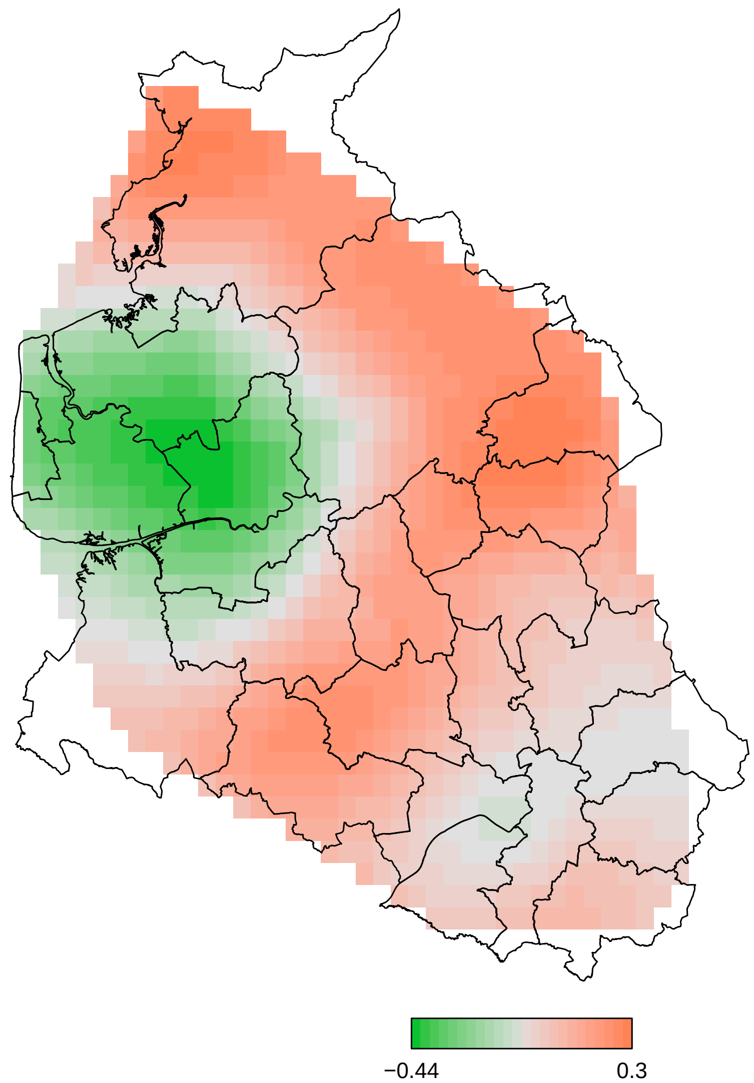
Effect of white blood cell count.



Effect of deprivation.



District-level analysis



Individual-level analysis

Structured Hazard Regression

- A general structured hazard regression model consists of an arbitrary combination of the following model terms:
  - Log baseline hazard $g_0(t) = \log(\lambda_0(t))$.
  - Time-varying effects $g_l(t)u_l$ of covariates u_l.
  - Nonparametric effects $f_j(x_j)$ of continuous covariates x_j.
  - Spatial effects $f_{spat}(s)$ of a spatial location variable s.
  - Interaction surfaces $f_{j,k}(x_j, x_k)$ of two continuous covariates.
  - Varying coefficient interactions $u_j f_k(x_k)$ or $u_j f_{spat}(s)$.
  - Frailty terms b_g (random intercept) or $x_j b_g$ (random slopes).
- All covariates are themselves allowed to be (piecewise constant) time-varying.

- Penalised splines for the baseline effect, time-varying effects, and nonparametric effects:
  - Approximate $f(x)$ (or $g(t)$) by a weighted sum of B-spline basis functions

$$f(x) = \sum \xi_j B_j(x).$$

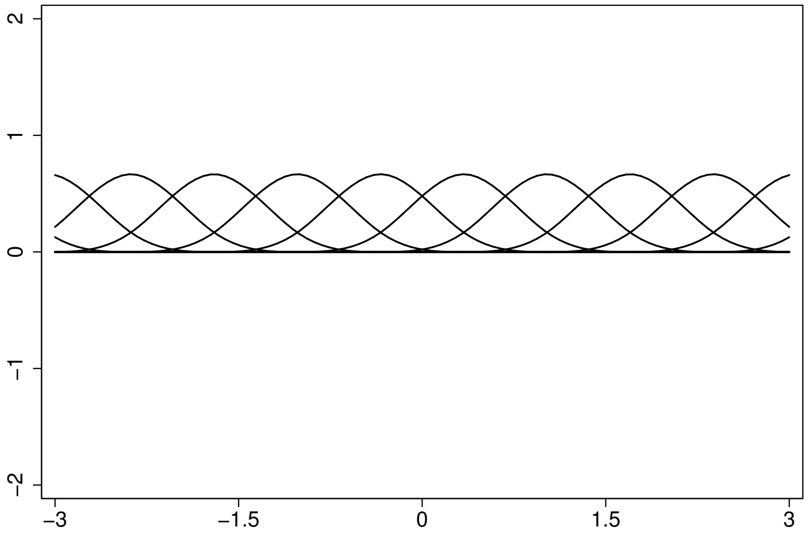
  - Employ a large number of basis functions to enable flexibility.
  - Penalise differences between parameters of adjacent basis functions to ensure smoothness:

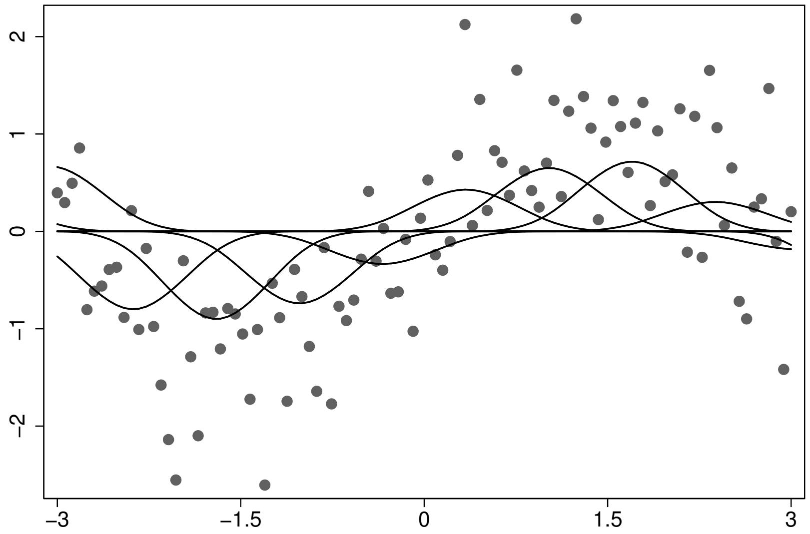
$$Pen(\xi|\tau^2) = \frac{1}{2\tau^2} \sum (\Delta_k \xi_j)^2.$$

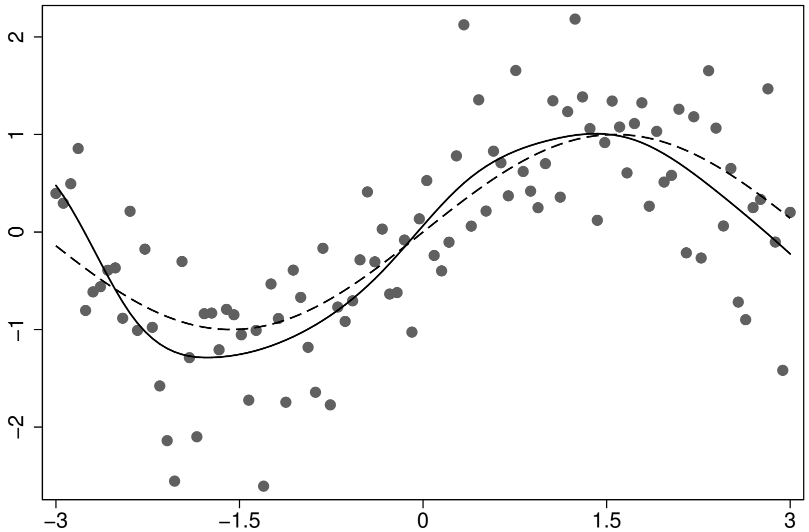
  - Bayesian interpretation: Assume a k-th order random walk prior for ξ_j, e.g.

$$\xi_j = \xi_{j-1} + u_j, \quad u_j \sim N(0, \tau^2) \qquad \text{(RW1)}.$$

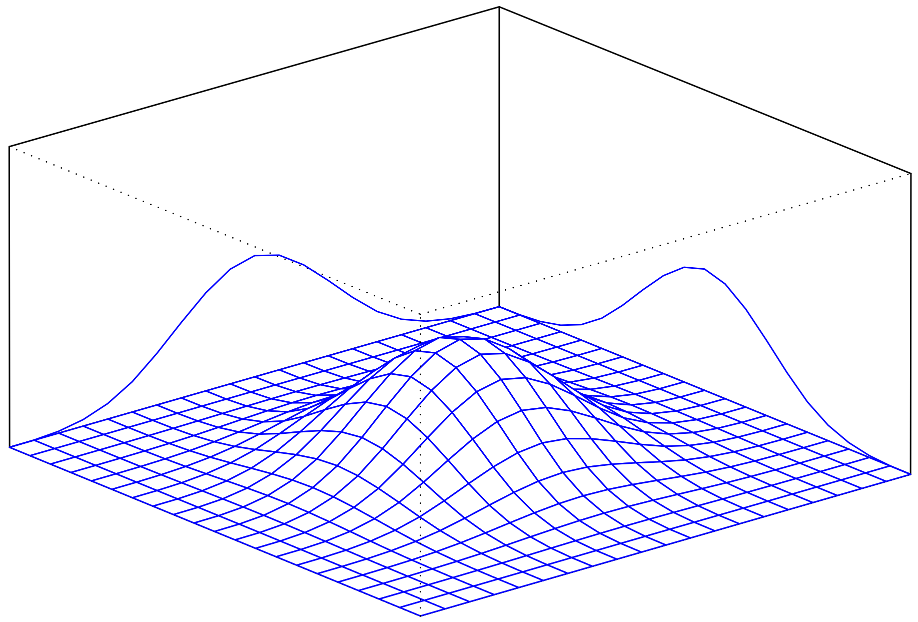
$$\xi_j = 2\xi_{j-1} - \xi_{j-2} + u_j, \quad u_j \sim N(0, \tau^2) \qquad \text{(RW2)}.$$

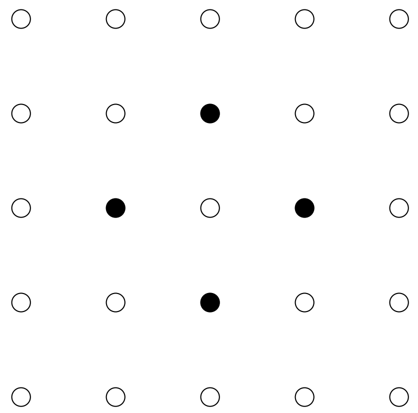

2
1
0
−1
−2
−3
−1.5
0
1.5
3


2
1
0
−1
−2
−3
−1.5
0
1.5
3


2
1
0
−1
−2
−3
−1.5
0
1.5
3

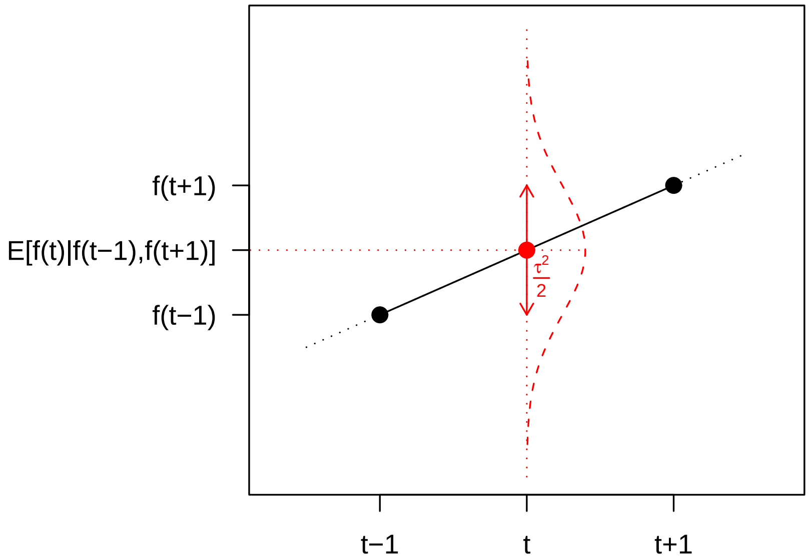
- Bivariate Tensor product P-splines for interaction surfaces:
  - Define bivariate basis functions (Tensor products of univariate basis functions).
  - Extend random walks on the line to random walks on a regular grid.

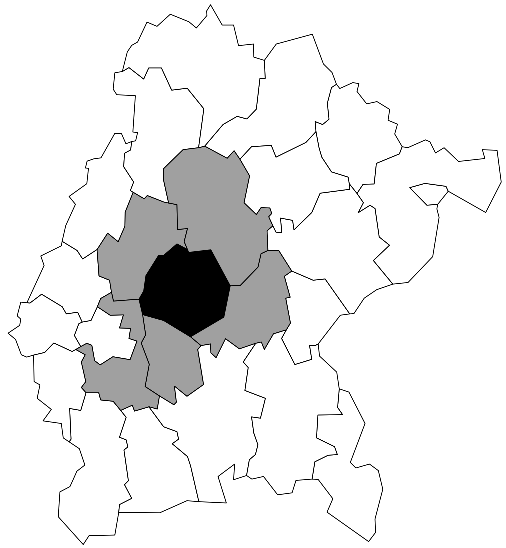




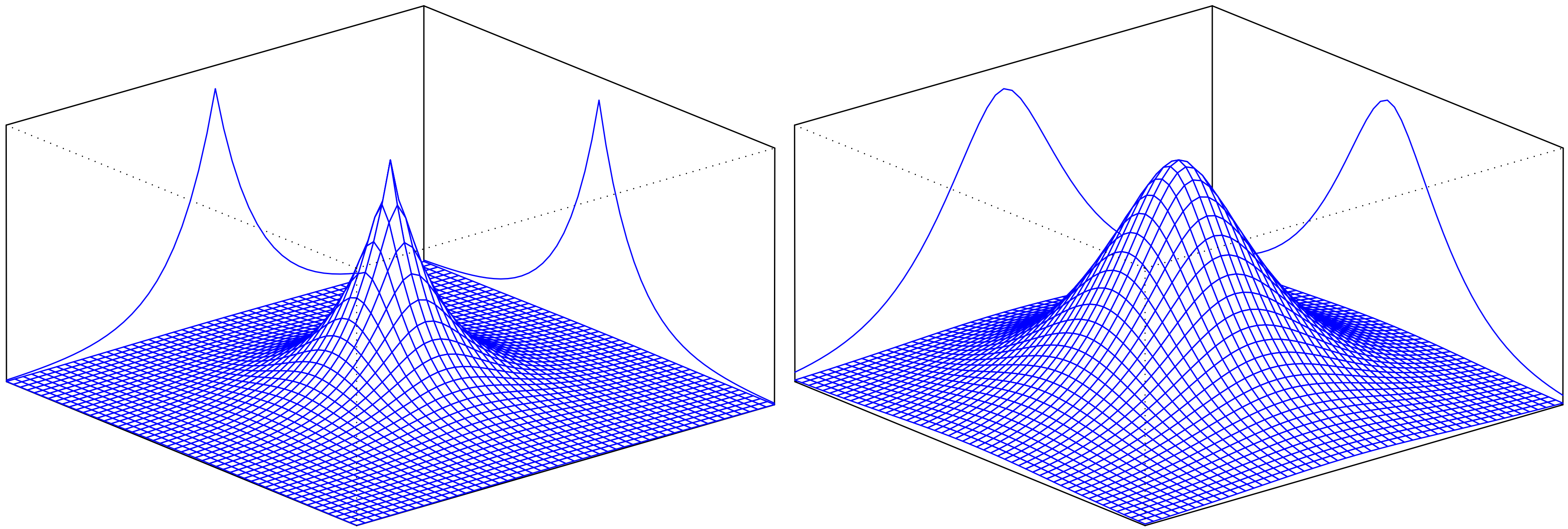
- Spatial effects for regional data $s \in \{1, \ldots, S\}$: (Intrinsic Gaussian) Markov random fields.
  - Bivariate extension of a first order random walk on the real line.
  - Define appropriate neighbourhoods for the regions.
  - Assume that the expected value of $f_{spat}(s) = \xi_s$ is the average of the function evaluations of adjacent sites:

$$\xi_s | \xi_{s'}, s' \neq s, \tau^2 \sim N\left(\frac{1}{N_s} \sum_{s' \in \partial_s} \xi_{s'}, \frac{\tau^2}{N_s} \right).$$





- Spatial effects for point-referenced data: Stationary Gaussian random fields.
  - Well-known as Kriging in the geostatistics literature.
  - Spatial effect follows a zero mean stationary Gaussian stochastic process.
  - Correlation of two arbitrary sites is defined by an intrinsic correlation function.
  - Can be interpreted as a basis function approach with radial basis functions.



- Cluster-specific frailty terms:
  - Account for unobserved heterogeneity.
  - Easiest case: i.i.d Gaussian frailty.
- All covariates in the discussed model terms are allowed to be piecewise constant time-varying.

Empirical Bayes Inference

- Generic representation of structured hazard regression models:

$$\lambda(t) = \exp\left[x(t)'\gamma + f_1(z_1(t)) + \ldots + f_p(z_p(t))\right]$$

- For example:

$f(z(t)) = g(t)$	$z(t) = t$	log-baseline effect,
$f(z(t)) = u(t)g(t)$	$z(t) = (u, t)$	time-varying effect of $u(t)$,
$f(z(t)) = f(x(t))$	$z(t) = x(t)$	smooth function of a continuous covariate $x(t)$,
$f(z(t)) = f_{spat}(s)$	$z(t) = s$	spatial effect,
$f(z(t)) = f(x_1(t), x_2(t))$	$z(t) = (x_1(t), x_2(t))$	interaction surface,
$f(z(t)) = b_g$	$z(t) = g$	i.i.d. frailty b_g, g is a grouping index.

- The generic representation facilitates description of inferential details.

- All vectors of function evaluations f_j can be expressed as

$$f_j = Z_j \xi_j$$

with design matrix Z_j, constructed from $z_j(t)$, and regression coefficients ξ_j.

- Generic form of the prior for ξ_j:

$$p(\xi_j|\tau_j^2) \propto (\tau_j^2)^{-\frac{k_j}{2}} \exp\left(-\frac{1}{2\tau_j^2}\xi_j' K_j \xi_j\right)$$

- $K_j \geq 0$ acts as a penalty matrix, $\mathrm{rank}(K_j) = k_j \leq d_j = \dim(\xi_j)$.

- $\tau_j^2 \geq 0$ can be interpreted as a variance or (inverse) smoothness parameter.

- Relation to penalized likelihood: Penalty terms

$$P_{\lambda_j}(\xi_j) = \log[p(\xi_j|\tau_j^2)] = -\frac{1}{2}\lambda_j \xi_j' K_j \xi_j, \qquad \lambda_j = \frac{1}{\tau_j^2}.$$

- Likelihood contributions for right- and uncensored survival times:

$$\lambda(T)^{\delta} \exp\left(-\int_0^T \lambda(t)dt\right),$$

where δ is the censoring indicator.

- Likelihood contributions for interval-censored observations:

$$\begin{aligned} P(T \in [T_{lower}, T_{upper}]) &= S(T_{lower}) - S(T_{upper}) \\ &= \exp\left[-\int_0^{T_{lower}} \lambda(t)dt\right] - \exp\left[-\int_0^{T_{upper}} \lambda(t)dt\right]. \end{aligned}$$

$\Rightarrow$ Derivatives of the log-likelihood become much more complicated for interval-censored survival times.

- In general, numerical integration has to be used to evaluate the cumulative hazard rate (e.g. the trapezoidal rule).
- Left truncation can easily be included.

- Principal idea of empirical Bayes estimation:
  - Differentiate between parameters of primary interest and hyperparameters.
  - Estimate the hyperparameters up-front from their marginal posterior.
  - Plug the resulting estimates back into the joint posterior and maximize with respect to the parameters of primary interest (yields posterior mode estimates).

- In structured hazard regression models:
  - regression coefficients are parameters of primary interest,
  - variance components are hyperparameters.

- Employ mixed model methodology to perform empirical Bayes inference: Consider ξ_j a correlated random effect with multivariate Gaussian distribution.

- Problem: In most cases partially improper random effects distribution $(k_j = \mathrm{rk}(K_j) < \dim(\xi_j) = d_j)$.

- Mixed model representation: Decompose

$$\xi_j = X_j\beta_j + Z_j b_j,$$

where

$$p(\beta_j) \propto const \qquad \text{and} \qquad b_j \sim N(0, \tau_j^2 I_{k_j}).$$

$\Rightarrow$ β_j is a fixed effect and b_j is an i.i.d. random effect.

- This yields the variance components model

$$\lambda(t; \cdot) = \exp\left[x'\beta + z'b\right],$$

where in turn

$$p(\beta) \propto const, \qquad b \sim N(0, Q),$$

and

$$Q = \text{blockdiag}(\tau_1^2 I, \ldots, \tau_p^2 I).$$

- Obtain empirical Bayes estimates / penalized likelihood estimates via iterating
  - Penalized maximum likelihood for the regression coefficients β and b.
  - Restricted Maximum / Marginal likelihood for the variance parameters in Q:

$$L^{marg}(Q) = \int L(\beta, b, Q)p(b)d\beta db \to \max_{Q}.$$

- Penalized score function and penalized Fisher information:

$$s_p(\beta, b) = \begin{pmatrix} \frac{\partial l(\beta, b)}{\partial \beta} \\ \frac{\partial l(\beta, b)}{\partial b} - Q^{-1}b \end{pmatrix}$$

$$F_p(\beta, b) = \begin{pmatrix} \frac{\partial^2 l(\beta, b)}{\partial \beta \partial \beta'} & \frac{\partial^2 l(\beta, b)}{\partial \beta \partial b'} \\ \frac{\partial^2 l(\beta, b)}{\partial b \partial \beta'} & \frac{\partial^2 l(\beta, b)}{\partial b \partial b'} - Q^{-1} \end{pmatrix}.$$

- Marginal likelihood estimation corresponds to REML estimation of variances in Gaussian mixed models.

- The marginal likelihood can not be derived analytically $\Rightarrow$ Apply a Laplace approximation.

- This yields the approximate marginal log-likelihood

$$l^{marg}(Q) \approx l(\hat{\beta}, \hat{b}) - \frac{1}{2}\log|Q| - \frac{1}{2}\hat{b}'Q^{-1}\hat{b} - \frac{1}{2}\log|F_p|,$$

  where F_p is the penalised Fisher information matrix.

- If both $l(\hat{\beta}, \hat{b})$ and $\hat{b}$ vary only slowly when changing the variance components we can further reduce the marginal log-likelihood to

$$l^{marg}(Q) \approx -\frac{1}{2}\log|Q| - \frac{1}{2}\log|F_p| - \frac{1}{2}b'Q^{-1}b,$$

  where b denotes a fixed value, e.g. a current estimate.

- This allows to device a Fisher Scoring algorithm based on matrix differentiation rules.

Software

- Implemented in BayesX, a software package for Bayesian inference in geoadditive and related models.

- Available from

`http://www.stat.uni-muenchen.de/~bayesx`

- More features:
  - Fully Bayesian inference based on MCMC in comparable model classes.
  - Univariate responses from exponential families (Gaussian, Binomial, Poisson, Negative Binomial, Gamma, . . .).
  - Categorical responses (multinomial regression models, cumulative models, sequential models).
- Latest development: Multi-state models.

Multi-State Models

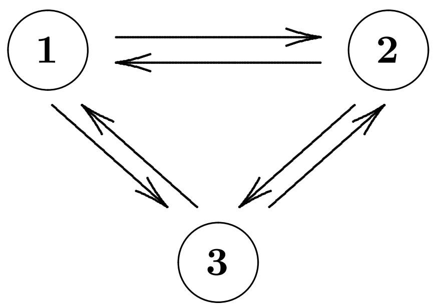
- Multi-state models form a general class for the description of the evolution of discrete phenomena in continuous time.

- We observe paths of a process

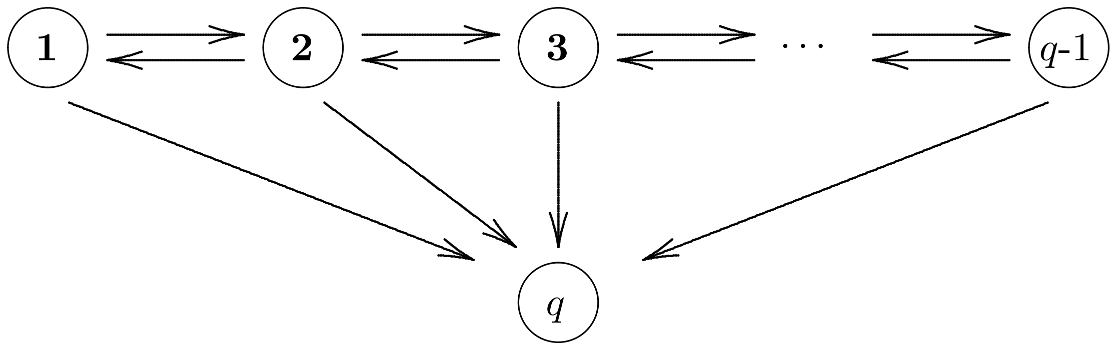
$$X = \{X(t), t \geq 0\} \quad \text{with} \quad X(t) \in \{1, \dots, q\}.$$

- Yields a similar data structure as for Markov processes.

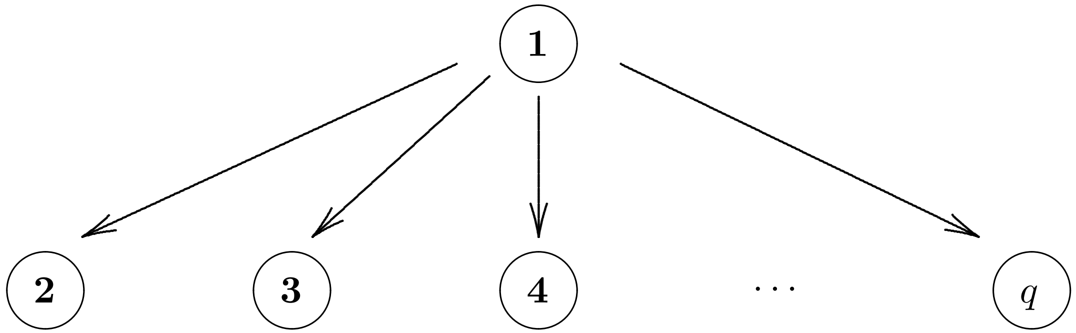
- Examples:
  - Recurrent events:



– Disease progression:



– Competing risks:



- (Homogenous) Markov processes can be compactly described in terms of the transition intensities

$$\lambda_{ij} = \lim_{\Delta t \to 0} \frac{P(X(t + \Delta t) = j | X(t) = i)}{\Delta t}$$

- Often not flexible enough in practice since
  - The transition intensities might vary over time.
  - The transition intensities might be related to covariates.
  - The Markov model implies independent and exponentially distributed waiting times.

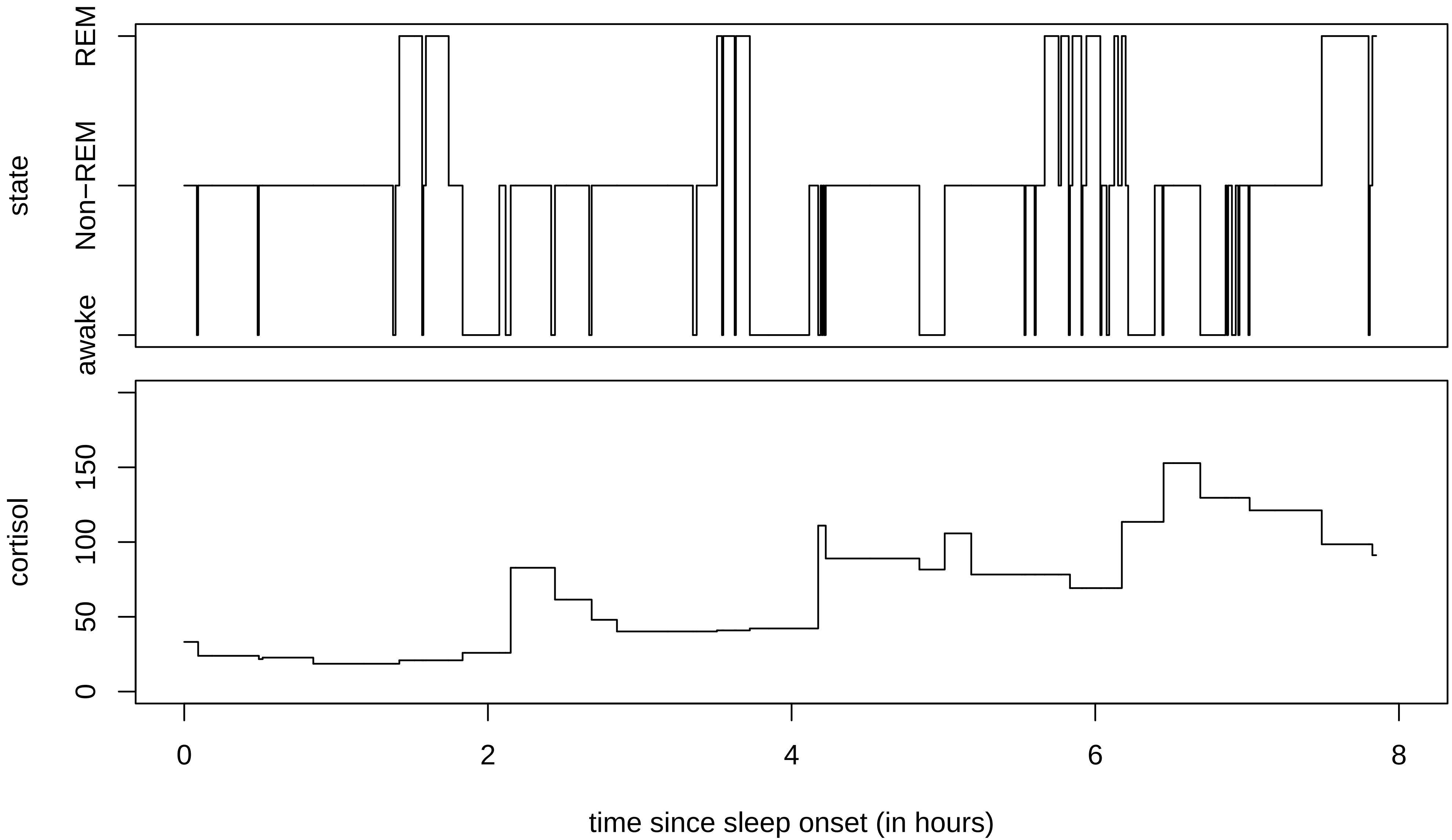
Human Sleep Data

- Human sleep can be considered an example of a recurrent event type multi-state model.

- State Space:

Awake	Phases of wakefulness
REM	Rapid eye movement phase (dream phase)
Non-REM	Non-REM phases (may be further differentiated)

- Aims of sleep research:
  - Describe the dynamics underlying the human sleep process.
  - Analyse associations between the sleep process and nocturnal hormonal secretion.
  - (Compare the sleep process of healthy and diseased persons.)


REM
Non−REM
awake
state
cortisol
0
50
100
150
0
2
4
6
8
time since sleep onset (in hours)

- Data generation:
  - Sleep recording based on electroencephalographic (EEG) measures every 30 seconds (afterwards classified into the three sleep stages).
  - Measurement of hormonal secretion based on blood samples taken every 10 minutes.
  - A training night familiarises the participants of the study with the experimental environment.

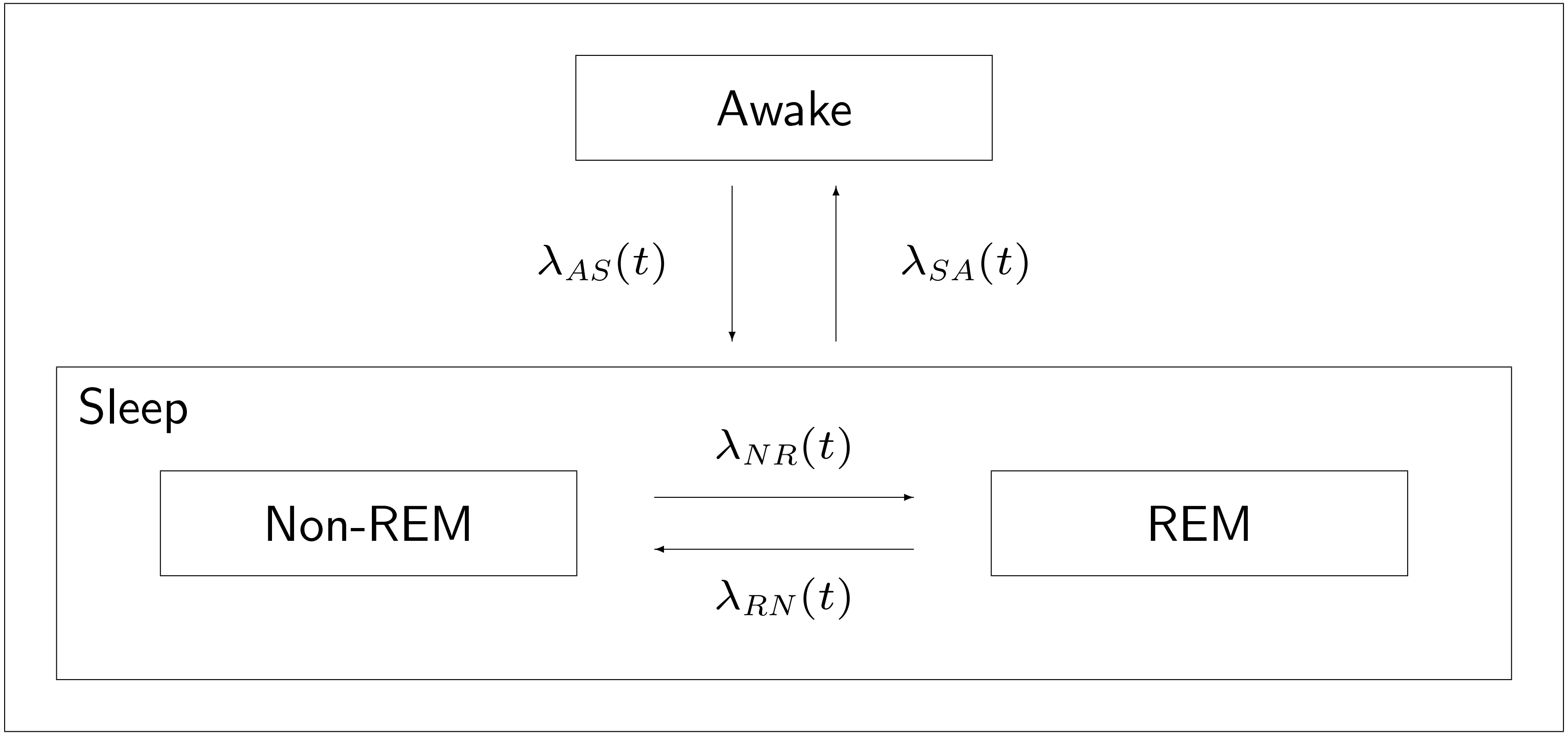
  ⇒ Sleep processes of 70 participants.

- Simple parametric approaches are not appropriate in this application due to
  - Changing dynamics of human sleep over night.
  - The time-varying influence of the hormonal concentration on the transition intensities.
  - Unobserved heterogeneity.

⇒ Model transition intensities nonparametrically.

Specification of Transition Intensities

- To reduce complexity, we consider a simplified transition space:



- Model specification:

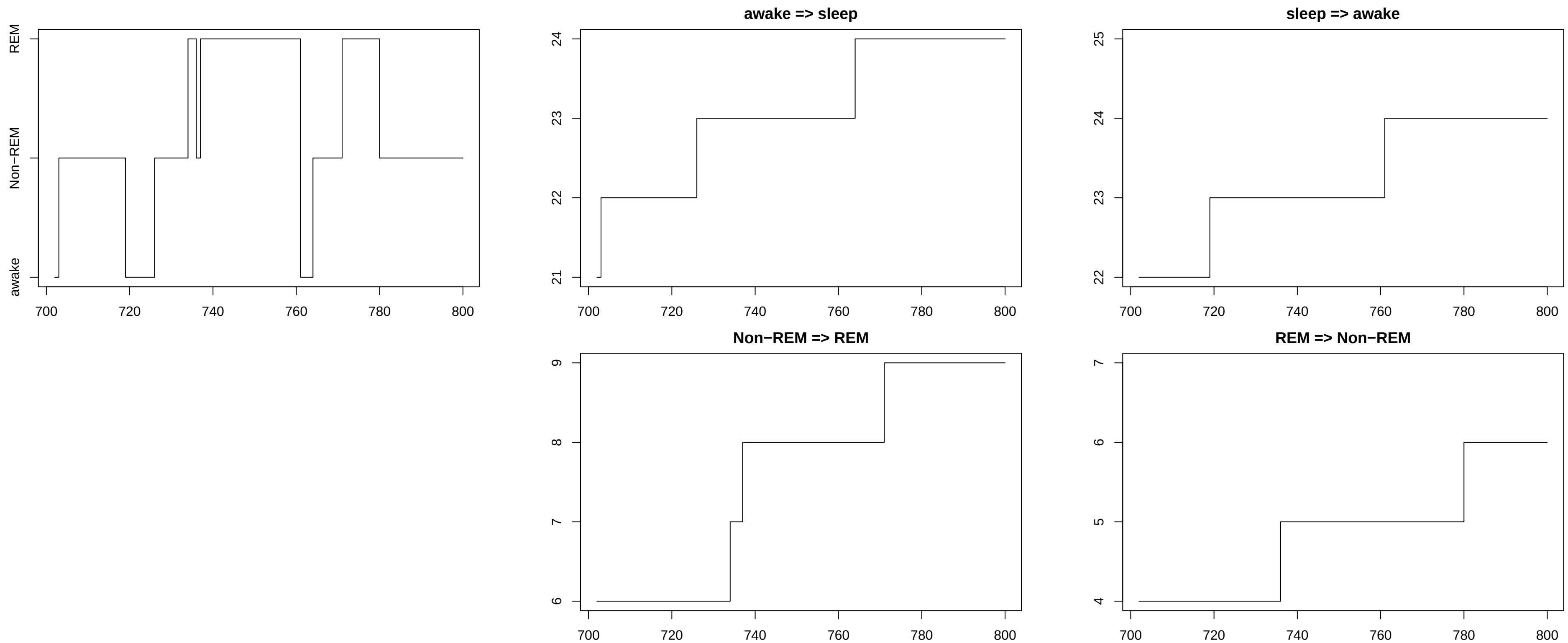
$$
\begin{aligned}
\lambda_{AS,i}(t) &= \exp\left[\gamma_0^{(AS)}(t) + b_i^{(AS)}\right] \\
\lambda_{SA,i}(t) &= \exp\left[\gamma_0^{(SA)}(t) + b_i^{(SA)}\right] \\
\lambda_{NR,i}(t) &= \exp\left[\gamma_0^{(NR)}(t) + c_i(t)\gamma_1^{(NR)}(t) + b_i^{(NR)}\right] \\
\lambda_{RN,i}(t) &= \exp\left[\gamma_0^{(RN)}(t) + c_i(t)\gamma_1^{(RN)}(t) + b_i^{(RN)}\right]
\end{aligned}
$$

where

$$
\begin{aligned}
c_i(t) &= \begin{cases} 1 & \text{cortisol} > 60 \text{ n mol/l at time } t \\ 0 & \text{cortisol} \leq 60 \text{ n mol/l at time } t, \end{cases} \\
b_i^{(j)} \sim N(0, \tau_j^2) &= \text{transition- and individual-specific frailty terms.}
\end{aligned}
$$

Counting Process Representation

- A multi-state model with k different types of transitions can be equivalently expressed in terms of k counting processes $N_h(t)$, $h = 1, \ldots, k$ counting these transitions.

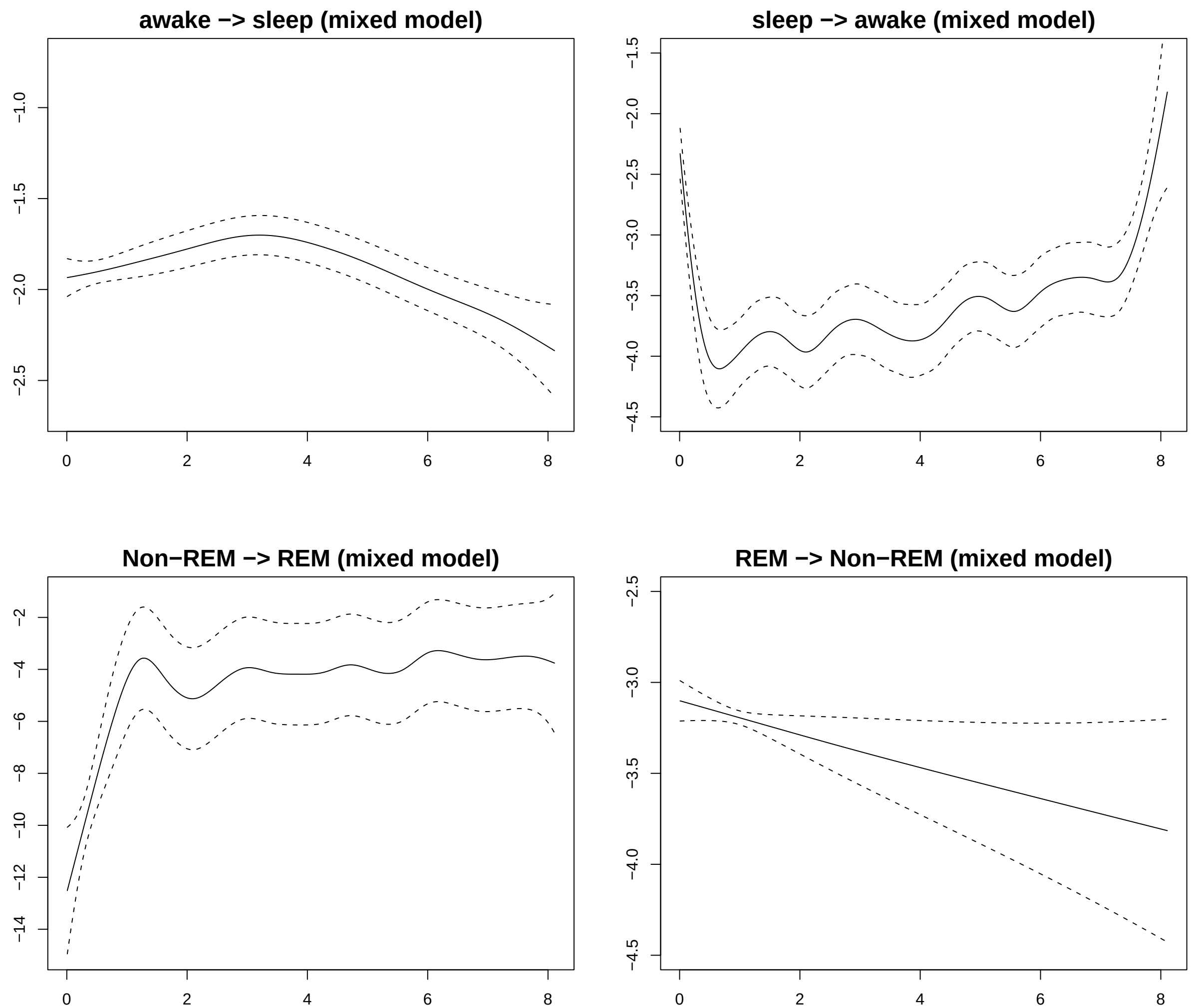


- From the counting process representation we can derive the likelihood contributions for individual i:

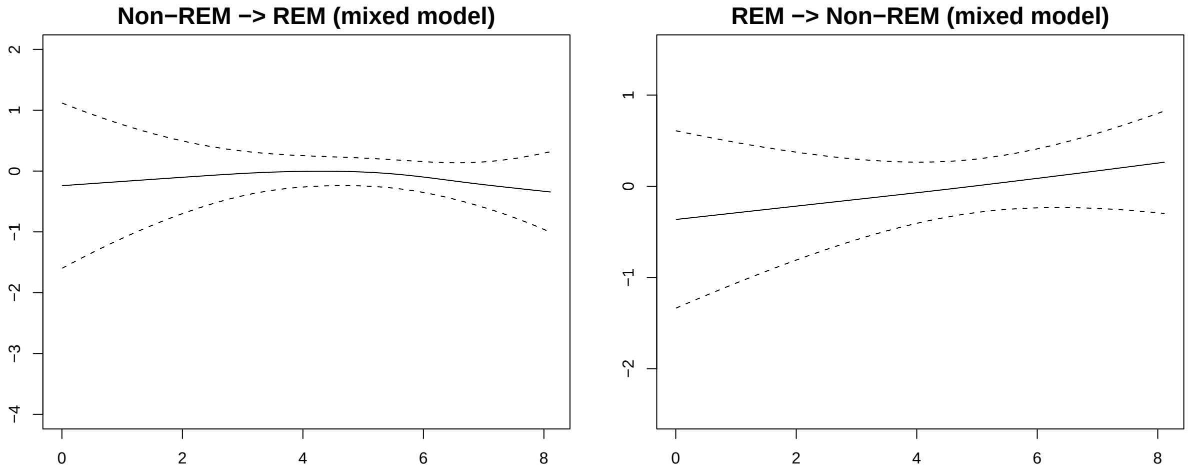
$$
\begin{aligned}
l_i &= \sum_{h=1}^{k} \left[\int_0^{T_i} \log(\lambda_{hi}(t)) dN_{hi}(t) - \int_0^{T_i} \lambda_{hi}(t) Y_{hi}(t) dt \right] \\
&= \sum_{j=1}^{n_i} \sum_{h=1}^{k} \left[\delta_{hi}(t_{ij}) \log(\lambda_{hi}(t_{ij})) - Y_{hi}(t_{ij}) \int_{t_{i,j-1}}^{t_{ij}} \lambda_{hi}(t) dt \right].
\end{aligned}
$$

k	number of possible transitions.
$N_{hi}(t)$	counting process for type h event and individual i.
$Y_{hi}(t)$	at risk indicator for type h event and individual i.
t_{ij}	event times of individual i.
n_i	number of events for individual i.
$\delta_{hi}(t_{ij})$	transition indicator for type h transition.

- Baseline effects:



- Time-varying effects for a high level of cortisol:



- Individual-specific variation is only detected for the transition between REM and Non-REM.

Conclusions

- Unified framework for general regression models describing the hazard rate of survival models.
- Empirical Bayes inference based on mixed model methodology.
- Extendable to models for transition intensities in multi state models.
- Future work:
  - More general censoring mechanisms for multi-state models.
  - Conditions for propriety of posteriors.
  - Joint modelling of covariates and duration times.

References

- Brezger, Kneib & Lang (2005). BayesX: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software*, **14** (11).

- Fahrmeir, Kneib & Lang (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14**, 731–761.

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

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

- Kneib & Hennerfeind (2006). Bayesian semiparametric multi-state models. In preparation.