Joint Dispersion Model with a Flexible Link

Rui Martins, ruimartins@egasmoniz.edu.pt
Centro de Investigação Interdisciplinar Egas Moniz (CiiEM),
Escola Superior de Saúde Egas Moniz (ESSEM), Portugal.

Objectives/Introduction

The objective is to model longitudinal and survival data jointly taking into account the dependence between the two responses in a real HIV/AIDS dataset. Linking parameters are time-dependent and specified using P-Splines, which allows for a considerable flexibility. In addition, standard deviation of the longitudinal measure is hierarchically modelled and considered as a covariate in the risk model.

We will connect the longitudinal and survival responses using a shared parameter approach inside a Bayesian framework.
For the longitudinal model it is usually assumed that all the residuals have a common variance. However, it is reasonable to suspect that residual variances for different subjects might differ. An hierarchical model is used to fit these patterns of variation by modelling the variance as a specified function of other variables. This will allow us to handle the heteroscedasticity (non constant variance).
Another important aspect of this work is the flexibility of the linking parameter(s). Traditionally they are considered time-independent. Instead we consider here a time-dependent specification allowing for a relationship varying in time between the longitudinal and survival models. In this context splines arise as a natural approach, in particular Penalized Cubic B-Splines, making the hazard model a regression model with time-varying coefficients.
We investigate the performance of the proposed method and apply it to analyze a real HIV/AIDS dataset.

Dataset

- **Data origin**: Brazilian database on HIV;
- **Period (years)**: 2002-2006;
- **Sample sizes**: $n = 500$ individuals
- **Response variables**: $y = \sqrt{\text{CD4}^{+}\text{T}}$ lymphocyte counts and survival time (years since the patient's entry in the study until the event)
- **Explanatory variables**: age (<50=0, ≥50=1); gender (Female=0, Male=1); PrevOI (previous opportunistic infection at study entry=1, no previous infection=0);
- 34 deaths. 88% of the patients were between 15 and 49 years old; 60% were males. The CD4 counts initial median was 245 cells/mm^3 (men - 226 cells/mm^3; women - 263 cells/mm^3).

References

[1] McLain, A.C., Lum, K.J. and Sundaram, R. (2012). A Joint Mixed Effects Dispersion Model for Menstrual Cycle Length and Time-to-Pregnancy. *Biometrics*, **68**:648–656.
[2] Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2006). Geoadditive Survival Models. *JASA*, **101**:1065–1075.
[3] Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *JCGS*, **13**:183–212.
[4] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *JMLR*, **11**(455):3571–3591.

Acknowledgments

The author is grateful to the organisers for the scholarship that allowed him to attend the ***Workshop on Flexible Models for Longitudinal and Survival Data with Applications in Biostatistics*** (27th-29th July, 2015).

Model

Consider the repeated measurements, $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$, and the observed (possibly right censored) time to event, T_i, for the ith individual, $i = 1, \ldots, N$. **Longitudinal data** is described by a **mixed effects dispersion model** [1],

$$y_{ij}|\boldsymbol{b}_i, \sigma_i^2 \sim \mathcal{N}(m_i(t_{ij}), \sigma_i^2), \quad j = 1, \ldots, n_i \tag{1}$$
$$m_i(t_{ij}) = \mathbf{x}_{1i}^{\top}\boldsymbol{\beta}_1 + \mathbf{w}_{1i}^{\top}\boldsymbol{b}_{1i}, \tag{2}$$
$$\sigma_i^2 = \sigma_0^2 \exp\{\, \mathbf{x}_{2i}^{\top}\boldsymbol{\beta}_2 + \mathbf{w}_{2i}^{\top}\boldsymbol{b}_{2i}\,\}, \tag{3}$$

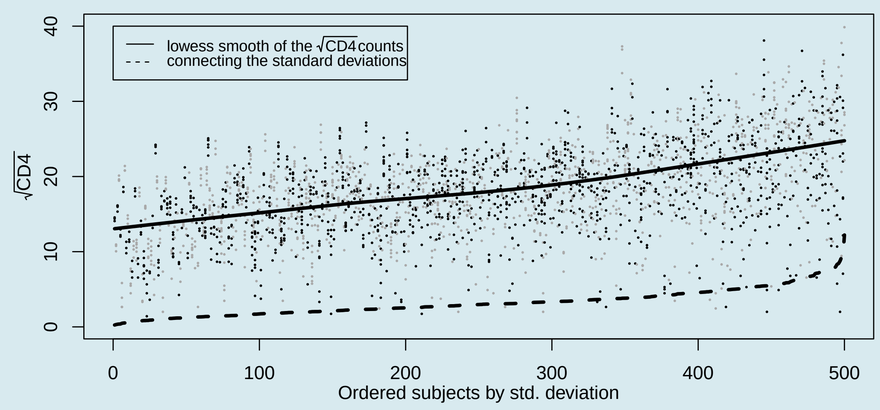
where $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$, $\mathbf{w}_{1i}$ and $\mathbf{w}_{2i}$ are appropriate subject-specific design vectors of covariates (possibly time-dependent); $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of population regression parameters; $(\boldsymbol{b}_{1i}^{\top}, \boldsymbol{b}_{2i}^{\top}) = \boldsymbol{b}_i \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$ are time-independent subject-specific random effects. To study the **association** between the longitudinal and the survival processes we consider the inclusion of some characteristics of the longitudinal trajectory into a **Relative Risk model**,

$$h_i(t \mid \mathbf{x}_{3i}, \boldsymbol{b}_i, \sigma_i) = h_0(t) \exp\{\, \mathbf{x}_{3i}^{\top}\boldsymbol{\beta}_3 + \mathcal{C}_i\{\boldsymbol{b}_i, \sigma_i; \boldsymbol{g}(t)\}\,\}, \tag{4}$$

where $\mathbf{x}_{3i}$ is a subject-specific design vector of baseline covariates. $\mathcal{C}_i\{.\}$ is a function specifying which components of the longitudinal process are related to $h_i(.)$. Finally, $\boldsymbol{g}(t)$ is an appropriate vector of smooth functions (approximated by **Penalized cubic B-Spline functions with 19 internal knots** over the domain of t [3]) representing the time-varying regression coefficients [2], which measure the effect of some particular characteristics of the longitudinal outcome to the hazard. Baseline hazard, $\boldsymbol{h_0(t)}$, can have a parametric form (*e.g.* **Weibull**) or be specified using a **P-Spline** or a **piecewise constant function**.

Reasoning behind the dispersion model

One of the common assumptions in longitudinal models is that the within-individual variances are homogeneous, although this assumption is not always satisfied. The plot of individual $\sqrt{\text{CD4}}$ values against the standard deviation (Figure 1) suggests considerable within-subject variance heterogeneity. Large values of the mean are associated with high variability.



In (3), σ_i^2 has a log-linear representation and is assumed to be an individual property allowing for heterogeneity in the variance trends (increasing or decreasing) among the individuals. Modelling it and identify covariates related to this discrepancies seems necessary. σ_0^2 acts as a "baseline" variance. $\mathbf{x}_{2i}$ and $\mathbf{w}_{2i}$ represent the variables influencing the within-subjects variance. In this way, one can examine whether contextual variables are related to the within-individual variance.

Application/Results

$m_i(t_{ij}) = [1\ t_{ij}\ \text{sex age PrevOI}]\boldsymbol{\beta}_1 + [1\ t_{ij}]\boldsymbol{b}_{1i}$;		$\mathbf{x}_{3i}^{\top}\boldsymbol{\beta}_3 = [\text{sex age PrevOI}]\boldsymbol{\beta}_3$		
		$\boldsymbol{h_0(t)}$		
$\boldsymbol{\sigma_i^2}$	$\boldsymbol{\mathcal{C}_i}$	**Weib**	**PSpline**	**Piecewise**
$\sigma_0^2 \exp\{b_{2i}\}$	$g_1(t)b_{1i,1} + g_2(t)b_{1i,2} + g_3(t)b_{2i}$	14700	12848	14483
$\sigma_0^2 \exp\{[\text{sex age PrevOI}]\boldsymbol{\beta}_2 + b_{2i}\}$		14671	12573	14317
σ_i^2	$g_1(t)b_{1i,1} + g_2(t)b_{1i,2} + g_3(t)\sigma_i$	13134	**12104♣**	12921
σ_0^2		13956	12887	13533
$\sigma_0^2 \exp\{b_{2i}\}$		14452	12917	13571
$\sigma_0^2 \exp\{[\text{sex age PrevOI}]\boldsymbol{\beta}_2 + b_{2i}\}$		14307	12605	13365
s.a.a.	$g_1(t)b_{1i,1} + g_2(t)b_{1i,2}$	16827	13553	14955
		17209	13688	15468
		15036	12544	14799
		15009	13334	14663

Table: **WAIC values** for the 24 joint models.

The most common approach to link the longitudinal and survival models is to consider shared random effects. Although, considering this particular dataset, the best fit is achieved when we share the individual random effects and the individual standard deviation is considered a covariate for the relative risks model (Model ♣). The results of this contribution need more studies to support them, namely concerning the linking structure, but they seem to be encouraging. In many HIV/AIDS studies, where investigators are interested in understanding the trends of the variability, because heteroscedasticity may be related to the survival time, having an individual estimate of the subject-residual variance can be a plus.