Smooth random effect distribution in joint models with an application to cardiomiopathy data.

Effetti casuali smooth nei modelli congiunti con applicazione a dati sulle cardiomiopatie.

Sara Viviani

Abstract Longitudinal studies often entail non Gaussian primary responses. If individuals prematurely leave the study, potential non-ignorability of the missingness process may occur. A joint model for the primary response and a time-to-event may represent an appealing tool to account for dependence between the two processes. As an extension to the generalized linear mixed joint models, recently proposed, and based on Gaussian latent effects, we assume that the random effects follow a smooth, P-spline based density. To estimate model parameters, we adopt a two-step conditional Newton-Raphson algorithm. Since the maximization of the penalized log-likelihood requires numerical integration over the random effect, which is often cumbersome, we opt for a Pseudo-Adaptive Gaussian quadrature rule to approximate the model likelihood. We discuss the proposed model by analyzing an original dataset on dilated cardiomyopathies.

Abstract *Gli studi longitudinali spesso sono soggetti al problema del dropout. Qualora il processo che genera i dati mancanti risulti non ignorabile, é necessario un modello congiunto che prenda in considerazione la dipendenza tra risposta longitudinale e tempo all'evento. Nel presente lavoro viene proposta un' estensione del modello lineare generalizzato congiunto, utilizzato per risposte longitudinali non Gaussiane e tempi di sopravvivenza, sotto l'assunzione di effetti casuali normali. Si assume che gli effetti casuali (condivisi) seguano una distribuzione smooth, modellata tramite P-splines. Per la stima dei parametri del modello, si utilizza un algoritmo Newton-Raphson condizionato in due passi. La massimizzazione della verosimiglianza penalizzata richiede un'integrazione numerica, risolta con una quadratura Pseudo adattiva. Il modello viene discusso tramite l'analisi di un dataset originale sulle cardiomiopatie dilatative.*

Key words: Joint models, MDCM data, smooth random effect distribution, mixture of P-splines.

Sara Viviani
ESS Department, FAO, Viale delle Terme di Caracalla, Roma, e-mail: sara.viviani@fao.org

1 Introduction

In this paper, we propose a new version of the joint model for a longitudinal response with distribution in the exponential family and a time-to-dropout. We relax the Gaussian assumption upon the random effect distribution and introduce a set of shared random effects following a smooth, spline-based distribution, see De Boor (1978) and Eilers and Marx (1996). Ghidey et al. (2004) propose a P-spline based distribution for random effects in linear mixed models. Using a similar perspective, Tsonaka et al. (2006) formulate a shared parameter model to handle non-monotone non-ignorable missingness with random effects distributed according to a mixture of Gaussian densities. Rather than using the discrete time representation, we opted to extend the work of Ghidey et al. (2004) to non Gaussian longitudinal outcomes with dropout, and propose an original rearrangement of the JM formulation of Wulfsohn and Tsiatis (1997). In this case, we assume that the *linear predictor* of the longitudinal process influences the hazard of dropout.
To estimate model parameters, we maximize a penalized likelihood function through a two-step, Newton-type, algorithm. As the observed data likelihood requires numerical integration over the random effects, in order to reduce the computational burden, we propose to use the pseudo-adaptive quadrature rule, see Rizopoulos (2012).
We discuss the behaviour of parameter estimates in the case of binary longitudinal responses with attrition, and apply the proposed method to an original dataset.

The rest of the paper is as follows. In Section 2 we briefly discuss the joint model structure in the case of non-Gaussian responses; in Section 3 a smooth version of random effect distribution is described. In Section 4 the model is applied to an original dataset on dilated cardiomyopathies. Section 5 gives some concluding remarks.

2 The model structure

In this Section, we discuss an extension of joint models to handle mixed models for non-Gaussian responses and smooth random effects in longitudinal studies with attrition. The proposed parametrization is based on joint modeling a longitudinal response, with distribution in the exponential family, and a time-to-event process, assuming that the linear predictor for the longitudinal model at time t influences the hazard of dropout at the same time.

Let us denote by $T_i = \min(T_i^*, C_i)$ the observed failure time for the ith individual, $i = 1, \ldots, n$, taken as the minimum between the true event time T_i^* and the censoring time C_i, which may correspond to the end of the follow-up. Let δ_i be the event indicator defined by $\delta_i = I(T_i^* \leq C_i)$, where $I(\cdot)$ is the indicator function. The outcome $Y_i(t)$ is repeatedly observed before T_i, and is missing for $t \geq T_i$; we assume that the longitudinal process is associated with T_i^*, i.e the *true* event time, but it is independent of the censoring time C_i. We assume $Y_i(t) \sim \mathrm{EF}(\theta_i(t))$, i.e. that $Y_i(t)$ is a random variable with distribution in the exponential family and natural parameter

denoted by $\theta_i(t)$.
The extension of JMs by Viviani et al. (2013) to non-Gaussian responses, from here on indicated as GLMJM, have adopted the following model structure:

$$\begin{cases} g\left(m_i(t)\right) = \eta_i(t) = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i(t) + \mathbf{b}_i^{\mathsf{T}}\mathbf{Z}_i(t) \\ h_i(t \mid M_i(t), \mathbf{W}_i) = h_0(t)\exp\{\boldsymbol{\gamma}^{\mathsf{T}}\mathbf{W}_i + \alpha m_i(t)\}. \end{cases} \tag{1}$$

Here, $\mathbf{X}_i(t)$ is a p-dimensional vector of predictors measured at occasion t and associated to the fixed effects $\boldsymbol{\beta}$, $\mathbf{Z}_i(t)$ is a q-dimensional vector of predictors measured at the same occasion and associated to the random effects $\mathbf{b}_i$, while $\mathbf{W}_i$ is a row vector of additional (time constant) covariates used to model the survival process, associated to the parameter vector $\boldsymbol{\gamma}$. In this context, $h_i(t)$ represents the hazard of dropout at time t, while $h_0(t)$ denotes the baseline risk function, which is typically left unspecified (Cox, 1972; Andersen and Gill, 1982). As suggested by Hsieh et al. (2006), adopting a nonparametric baseline hazard in the JM framework may lead to underestimate the parameter standard errors; for this reason, we consider a Weibull baseline function, i.e. a baseline risk parametrized as $h_0(t) = \xi t^{\xi-1}$. The term $M_i(t)$ in the conditioning denotes the history of the true, unobserved, longitudinal process up to t, while $m_i(t) = g^{-1}(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i(t) + \mathbf{b}_i^{\mathsf{T}}\mathbf{Z}_i(t))$ is the expected value of the longitudinal response at time t, and $g(\cdot)$ is the link function.

We propose to modify model (6) as follows:

$$\begin{cases} g\left(m_i(t)\right) = \eta_i(t) = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i(t) + \mathbf{b}_i^{\mathsf{T}}\mathbf{Z}_i(t) \\ h_i(t \mid M_i(t), \mathbf{W}_i) = h_0(t)\exp\{\boldsymbol{\gamma}^{\mathsf{T}}\mathbf{W}_i + \alpha \eta_i(t)\}. \end{cases} \tag{2}$$

When compared to the standard GLMJM structure, where we assume the expected value of the response is shared by the longitudinal and the survival processes, this approach allows to prevent some computational issues. In fact, the expected value is, in most of the cases, on the exponential scale and, in rare empirical cases, may take infinite values. Moreover, this may lead to an always increasing risk for each value of the fitted mean when compared to the baseline risk.
The parameter α can be seen as non-ignorability parameter, since it gives an estimate of the degree of association between the dropout and the longitudinal processes.

The survival function corresponding to model (2) is

$$\begin{aligned} S(T_i \mid M_i(T_i), \mathbf{W}_i; \boldsymbol{\gamma}, \alpha, \xi) &= \exp\left\{-\int_0^{T_i} h_i(s)ds\right\} \\ &= \exp\left\{-\int_0^{T_i} \xi s^{\xi-1}\exp\{\boldsymbol{\gamma}^{\mathsf{T}}\mathbf{W}_i + \alpha\eta_i(s)\}ds\right\}. \end{aligned} \tag{3}$$

Since equation (3) involves an integral over the time which is not analytically tractable, we approximate it by Gauss-Kronrod numerical integration method, see Kronrod (1964).

3 Random Effect distribution

To consider random effects with flexible distribution, we follow the approach proposed by Ghidey et al. (2004). We consider a zero mean bivariate random effect $\mathbf{b}_i = (b_{i1}, b_{i2})$ with covariance matrix $\mathbf{D}$, and adopt the variable change $\mathbf{b}_i = \mathbf{D}^{1/2}\mathbf{b}_i^*$, where $\mathbf{b}_i^*$ are standardized random coefficients. From here on, we will denote by $\mathbf{C} = \mathbf{D}^{1/2}$. We assume that $\mathbf{b}_i^*$ take values on a square $[-m,m] \times [-m,m]$, and rather than adopting a parametric distribution for $\mathbf{b}_i$, we define a semi-parametric distribution as a finite mixture of bivariate Gaussian densities. As it is well known, this distribution may be used to fit, mimicking a Kernel-based centers approach, a wide variety of continuous distributions. We assume that the means are component-specific, say $\boldsymbol{\mu}_{1j}$, $j = 1,\dots,J$, and $\boldsymbol{\mu}_{2l}$, $l = 1,\dots,L$. To allow for varying degrees of distance from the homogeneous model, we allow for different number of components, J and L, in each dimension. We fix profile-specific standard deviations, τ_1 and τ_2 for the first and the second profile, respectively, and, following Ghidey et al. (2004), we define them as functions of corresponding means:

$$\begin{cases} \tau_1 = \frac{2}{3}(\mu_{1j} - \mu_{1,j-1}), & j = 1,\dots,J \\ \tau_2 = \frac{2}{3}(\mu_{2l} - \mu_{2,l-1}), & l = 1,\dots,L. \end{cases}$$

According to these model assumptions, the random effect distribution is a mixture of Gaussian densities:

$$f(\mathbf{b}_i) = \sum_{j=1}^{J}\sum_{l=1}^{L} c_{jl} N(\mathbf{C}\mu_{jl}, \mathbf{C}D_s\mathbf{C}'), \tag{4}$$

where the weights $c_{jl} = \exp\{a_{jl}\}/\sum_{k=1}^{J}\sum_{m=1}^{L}\exp\{a_{km}\}$ are transformed spline coefficients, under the sum constraint $\sum_{j=1}^{J}\sum_{l=1}^{L} c_{jl} = 1$, and $\mathbf{D}_s = diag(\tau_1^2, \tau_2^2)$.

The smooth density in (4) is a mixture with $(J \times L)$ components, where component-specific means and variances are fixed in the grid $[-m,m] \times [-m,m]$.

Parameter estimation, as usual in nonparametric models, is based on specifying a penalized likelihood function, see among others Eilers and Marx (1996). The penalty terms are typically introduced to balance smoothness, pursued through a high number of knots, and robustness, in order to avoid overfitting while maintaining a good fit to the adopted distribution. Let us define the overall parameter vector by $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \xi, \boldsymbol{\gamma}, \alpha, \mathbf{a})$, where $\mathbf{a} = \{a_{jl}\}$, $j = 1,\dots,J$, $l = 1,\dots,L$. The penalized log-likelihood for the JM follows:

$$\begin{aligned} \ell_p(\boldsymbol{\theta}) = & \sum_{j=1}^{J}\sum_{l=1}^{L} \frac{\exp\{a_{jl}\}}{\sum_{k=1}^{J}\sum_{m=1}^{L}\exp\{a_{km}\}} \times \\ & \int f(Y_i|b_{i1},b_{i2}) f(T_i,\delta_i|b_{i1},b_{i2}) f(\mathbf{b}_i) db_{i1} b_{i2} \\ & - \frac{\lambda_1}{2}\sum_{j=1}^{J}\sum_{l=1}^{L}\Delta_1^k a_{jl} - \frac{\lambda_2}{2}\sum_{j=1}^{J}\sum_{l=1}^{L}\Delta_2^k a_{jl}, \end{aligned} \tag{5}$$

where λ_1 and λ_2 are the penalty terms, and Δ_q^k, $q = 1,2$, are k-th order difference operators representing the k-th order discrete derivatives in Tibshirani (2013). We follow the suggestion of Ghidey et al. (2004), and choose $k = 3$, since for a large number of basis, the smooth density converges in distribution to a Gaussian as $\boldsymbol{\lambda} \to \infty$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$.
The integral over the random effect space in (5) needs to be numerically evaluated $(J \times L)$ times, each one weighted by the corresponding c_{jl}. The standard procedure in JMs is to approximate the integral through the following finite sum:

$$\int f(Y_i|b_{i1}, b_{i2}) f(T_i, \delta_i|b_{i1}, b_{i2}) f(\mathbf{b}_i) db_{i1} b_{i2} =$$
$$\sum_{(-m,m)\times(-m,m)} f(Y_i|b_{i1} = b_{t_1}, b_{i2} = b_{t_2}) f(T_i, \delta_i|b_{i1} = b_{t_1}, b_{i2} = b_{t_2}) f(b_{t_1}, b_{t_2}),$$

where $(t_1, t_2) \in \{1, \ldots, q\}$, q is the number of chosen discrete points where the integral is evaluated, and $(b_{t_1}, b_{t_2}) \in (-m,m) \times (-m,m)$. Since the integrand function needs to be adapted to the random effect posterior distribution, we opt for the Pseudo-Adaptive Gaussian (PA) rule, see Rizopoulos (2012) and Viviani et al. (2013), which looks for a better approximation with a substantial lowering of the computational complexity. In fact, the PA method computes the integrand by scaling and centring the random effect distribution only once, at the beginning of the optimization algorithm, leading to a faster fitting.
Estimation algorithm (not presented here) is a two-step conditional maximization Newton-Raphson procedure. Penalty terms are chosen through an ad hoc AIC formulation.

4 Dilated cardiomyopathy data

In this Section, an application to an original data set about patients with dilated cardiomyopathy is presented.

Dilation of the left ventricular is known to lead to hearth failure, and in many cases an ethiological basis cannot be identified (Merlo et al., 2011). The goal of this study is to compare patients with mild dilation of the left ventricular (Mildly Dilated CardioMyopathy or MDCM) with respect to patients with a general dilation of unrecognized ethiology (Idiopatic Dilated CardioMyopathy or IDCM). MDCM patients are generally believed to be at a slightly lower risk (e.g., Keren et al. (1990)), but the physiological reasons are still unrecognized.

Data refers to a study conducted in the cardiovascular department of "Ospedali Riuniti" in Trieste, Italy. The patients ($n = 659$) were enrolled at the first treatment for dilated cardiomyopathy, and then followed-up after 6 months, 1, 2, 3, 4, 6 and 10 years. At the end of the study, 25% of the patients had complete records, with a total of 212 events (32%).

The longitudinal outcome of interest is the New York Hearth Association (NYHA) functional classes, which places patients in one of four categories based on how much they are limited during physical activity related to the heart failure. Dropout occurs due to cardiovascular death, and it can be easily expected that the two processes are related as patients with a higher NYHA class are at higher risk of death.

We model the longitudinal outcome conditionally on an intercept, the visit occasion, the age at baseline and the indicator of MDCM. As for the survival model, we let the hazard of dropout depend on the gender indicator (1 for males). The joint model can be written as follows:

$$\begin{cases} g\left(m_i(t)\right) = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_i + \beta_2 MCDM_i + \beta_3 Age_i \\ h_i(t \mid M_i(t), \mathbf{W}_i) = h_0(t)\exp\{\gamma_1 Gender_i + \alpha m_i(t)\}. \end{cases} \tag{6}$$

A shared subject-specific random intercept and a random slope (referred to the visit occasion t_i) are included. The random effects are modeled through a mixture of Gaussian bases, weighted by transformed P-spline coefficients. For what concerns the spline settings, we fix $J = L = 5$ over the square $[-4,4] \times [-4,4]$, while the penalty terms are set to $\lambda_1 = 3$ and $\lambda_2 = 4$, as suggested by the AIC computation (see the previous Section).

Table 1 shows parameter estimates for the longitudinal and the survival submodels, comparing a MAR model, a GLMJM and the proposed smooth version.

Longitudinal			
	MAR	GLMJM	GLMJM smooth
Intercept	-0.600*	-0.593*	-0.566 *
	(0.1304)	*(0.0041)*	*(0.0384)*
t_i	-0.001	-0.001*	-0.002*
	(0.0006)	*(0.0001)*	*(0.0002)*
MDCM	-0.320*	-0.316*	-0.271*
	(0.0777)	*(0.0013)*	*(0.0788)*
Age	0.016*	0.015*	0.004 *
	(0.0027)	*(0.0011)*	*(0.0002)*
$\hat{\alpha}$		-0.196*	-1.544*
		(0.0012)	*(0.0003)*
Survival			
Gender	0.512*	0.496*	0.799*
	(0.1724)	(0.0001)	(0.0003)
log-lik	-730.5	-338.024	-211.005

Table 1 MDCM data: longitudinal and survival parameter estimates for the MAR, GLMJM and smooth random effect joint model (standard errors in brackets). The asterisks stand for significant coefficients.

Parameter estimates are comparable across model specifications. However, the smooth GLMJM results in a higher log-likelihood.

5 Discussion

In this paper, we have focused on mixed joint models for discrete outcomes by considering non Gaussian random effects. In particular, we have assumed that the latent effects follow a flexible smooth density, defined as a mixture of Gaussian densities based on P-splines, starting from the parametric approach of GLMJM by Viviani et al. (2013). Gaussianity of the random effects is a standard assumption for joint models that may be inappropriate to describe individual-specific sources of heterogeneity, and, at the same time, the dependence between the longitudinal and the survival processes.

An additional change to the GLMJM has been introduced: the longitudinal and the survival processes have been assumed to share the model response linear predictor, rather than the expected value of a generalized linear mixed model, to avoid computational issues due to the exponential scale of the expected value.

The application to the MDCM data set suggests that the smooth GLMJM presents a more refined prediction, when compared to the MAR and the standard GLMJM.

References

Alfó, M. and Aitkin, M. (2000). Random coefficient model for binary longitudinal responses with attrition. *Statistics and Computing*, 10:279–287.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, 10:1100–1120.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220.

De Boor, C. (1978). *A practical guide to splines.* Applied Mathematical Sciences.

Eilers, P. and Marx, D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11:89–121.

Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, 60:945–953.

Hsieh, F., Tseng, Y.-K., and Wang, J.-K. (2006). Joint modeling of survival and longitudinal data:likelihood approach revisited. *Biometrics*, 62:1037–1043.

Keren, A., Gottlieb, S., Tzivoni, D., Stern, S., Yarom, R., Billingham, M., and Popp, R. (1990). Mildly dilated congestive cardiomyopathy. Use of prospective diagnostic criteria and description of the clinical course without heart transplantation. *Circulation*, 81:506–517.

Kronrod, A. S. (1964). Doklady akad (in russian). *Nauk SSSR*, 154:283–286.

Merlo, M., Pyxaras, S., Pinamonti, B., Barbati, G., Di Lenarda, A., and Sinagra, G. (2011). Prevalence and prognostic significance of left ventricular reverse remodeling in dilated cardiomyopathy receiving tailored medical treatment. *Journal of the American College of Cardiology*, 57:1468–1476.

Rizopoulos, D. (2012). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics and Data Analysis*, 56:491–501.

Tibshirani, R. J. (2013). Adaptive piecewise polynomial estimation via trend filtering. *arXiv:1304.2986v1*.

Tsonaka, R., Verbeke, G., and Lesaffre, E. (2006). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics*, 65:81–87.

Viviani, S., Alfó, M., and Rizopoulos, D. (2013). Generalized linear mixed joint model for longitudinal and survival outcomes. *Statistics and Computing*, pages DOI 10.1007/s11222–013–9378–4.

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339.