A multivariate extension of the joint models

Un’estensione multivariata dei modelli congiunti

Marcella Mazzoleni and Mariangela Zenga

Abstract The joint models analyse the effect of longitudinal covariates onto the risk of an event. They are composed of two sub-models, the longitudinal and the survival sub-model. For the longitudinal sub-model a multivariate mixed model can be proposed. Whereas for the survival sub-model, a Cox proportional hazards model is proposed, considering jointly the influence of more than one longitudinal covariate onto the risk of the event. The purpose of the work is to extend an estimation method based on a joint likelihood formulation to the case in which the longitudinal sub-model is multivariate through the implementation of an Expectation-Maximisation (EM) algorithm.

Abstract *I modelli congiunti analizzano l’effetto delle covariate longitudinali sul rischio di un evento. Sono composti da due sotto-modelli, quello longitudinale e quello di sopravvivenza. Per il sotto-modello longitudinale si puó proporre un modello misto multivariato, mentre per quello di sopravvivenza viene proposto un modello a rischi proporzionali di Cox, dove le covariate longitudinali influenzano congiuntamente il rischio dell’evento. Lo scopo del lavoro é di estendere un metodo di stima basato sulla massimizzazione della verosimiglianza congiunta al caso in cui il sotto-modello longitudinale è multivariato attraverso l’implementazione di un algoritmo Expectation-Maximization (EM).*

Key words: Joint models, Multivariate Mixed Model, EM Algorithm, Joint Likelihood

Marcella Mazzoleni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milano, Italy e-mail: marcella.mazzoleni@unimib.it

Mariangela Zenga
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milano, Italy e-mail: mariangela.zenga@unimib.it

1 Introduction

The joint models analyse the effect of longitudinal covariates onto the risk of an event. They are composed of two sub-models, the longitudinal and the survival sub-model. For the longitudinal sub-model a multivariate mixed model can be proposed, considering fixed and random effects. Whereas for the survival sub-model, a Cox proportional hazards model is usually proposed, considering jointly the influence of more than one longitudinal covariate onto the risk of the event.

The joint models are often used in medical research because in clinical trails the aim is to analyse two subgroups, placebo and treated, in order to study the longitudinal covariates that could influence the survival time.

The first authors that extended the joint model to the case in which the longitudinal sub-model is multivariate are Xu and Zeger [10]. A Markov chain Monte Carlo algorithm was used to estimate parameters in the model, extending the univariate estimation method introduced by Xu and Zeger [11] and Faucett and Thomas [2]. The authors applied the model and the estimation method to the Schizophrenia trial data of risperidone. Albert and Shih [1] proposed a regression calibration approach for jointly modelling multiple longitudinal measurements and discrete time-to-event data. The authors proposed a two-stage regression calibration approach. Recently Hickey et al. [5] proposed a interesting review of all the model and estimation methods for the joint modelling of time-to-event and multivariate longitudinal outcomes. Despite developments, software to estimate the parameters of these model is still lacking. For this reason, Hickey et al. [4] implemented a new package in software R, the joineRML package. This package fits the joint model proposed by Henderson et al. [3], extended to the case of multiple continuous longitudinal measures. The association between time-to-event and longitudinal data is captured by a multivariate latent Gaussian process. The parameter are estimated using a Monte Carlo Expectation Maximization algorithm.

The purpose of the paper is to extend an estimation method based on a joint likelihood formulation used in the univariate case [6] to the case in which the longitudinal sub-model is multivariate. The parameters are estimated maximising the likelihood function, using an Expectation-Maximisation (EM) algorithm. In addition, in the M-step a one-step Newton-Raphson update is used, as for some parameters estimators, it is not possible to obtain closed-form expression. In addition, a Gauss-Hermite approximation is applied for some of the integrals involved.

2 Model and estimation method

The longitudinal and the survival sub-models compose the joint models. Concerning the survival sub-model, in this paper a proportional hazard model is used, which is defined as a function of the $m_{iq}(t)$ that denotes the true and unobserved value of the longitudinal covariate q for subject i:

$$h_i(t|M_i(t),\omega_i) = h_0(t)\exp\left[\gamma'\omega_i + \sum_q \alpha_q m_{iq}(t)\right] \tag{1}$$

where $M_i(t) = \{m_{iq}(s), 0 \leq s < t, \forall q = 1,\dots,Q\}$ indicates the history of the true unobserved longitudinal processes up to time t, α_q quantifies the effect of the longitudinal outcome q onto the risk of an event, $h_0(t)$ indicates the baseline hazard function, and ω_i are the covariates that influence the risk of the event with coefficient γ. Concerning the longitudinal sub-model, a linear multivariate mixed model is proposed:

$$y_{iq}(t) = m_{iq}(t) + \varepsilon_{iq}(t) \tag{2}$$

where q is the longitudinal variable index, $y_{iq}(t)$ is composed by the $m_{iq}(t) = x'_{iq}(t)\beta_q + z'_{iq}(t)b_{iq}$ and by a random error term $\varepsilon_{iq}(t) \sim N(0,\sigma^2)$, and β_q are the fixed effects for $x_{iq}(t)$, while b_{iq} are the random effects for $z_{iq}(t)$. In addition, $b'_i = (b'_{1q},...,b'_{iQ}) \sim N(0,D)$ and $b_{1q},...,b_{nQ}$ and $\varepsilon_{1q},...,\varepsilon_{nQ}$ are independent.

There are two classes of estimation method, the two-stage approach and the joint likelihood formulation. The two-stage approach is biased but less computationally demanding, while the joint likelihood is more efficient but computationally slower. The two-stage approach is based on two steps. In the first one the random effects are estimated using a least-squares approach, while in the second step the estimates previously found are used to impute appropriate values of $m_{iq}(t)$ that are substituted in the classical partial likelihood of the Cox model. The joint likelihood could be based on maximum likelihood, a Bayesian estimation of joint models using MCMC, or some hypothesis concerning the normal distribution of random effects or of covariates. Rizopoulos [6] proposed a new method of estimation based on the joint likelihood formulation, maximising the log-likelihood function through the Expectation-Maximisation (EM) and the Newton-Raphson algorithm.

The aim of the paper is to extend this method of estimation [6], to the case in which the longitudinal sub-model is multivariate. Starting from the classical log-likelihood equation, for each subject i it can be defined as:

$$\begin{aligned}\log p(T_i,\delta_i,y_i;\Theta) &= \log\int p(T_i,\delta_i,y_i,b_i;\Theta)db_i \\ &= \log\int p(T_i,\delta_i|b_i;\theta_t,\beta)\left\{\prod_q p(y_{iq}|b_{iq};\theta_y)\right\}p(b_i,\theta_b)db_i\end{aligned}$$

where $\Theta = (\theta'_t,\theta'_y,\theta'_b)'$ denotes the full parameter vector, with θ_t denoting the parameters for the event time outcome, θ_y the parameters for the longitudinal outcomes, and θ_b the unique parameters of the random-effects covariance matrix. In formula, $\theta_y = [\beta',\sigma^2]$ where $\beta = [\beta_1,...,\beta_q,...,\beta_Q]$ and $\sigma^2 = [\sigma_1^2,,...,\sigma_q^2,...,\sigma_Q^2]$; $\theta_t = [\gamma',\alpha_1,...,\alpha_q,...,\alpha_Q,\theta_{h_0}]$ where θ_{h_0} is used in the case in which the baseline hazard is parametric; and $\theta_b = [vech(D)]$. It is possible to separate the log-likelihood in three parts, where each part is related only to a part of the vector of parameters involved.

For maximising the log-likelihood function the Expectation-Maximisation (EM) al-

gorithm is used where the random effects are treated as ”missing data”. Accordingly, for the E-step the expected value of the complete data log-likelihood function considering the random effects as the missing data is considered. A numerical integration procedures must be employed as an integral with respect to the random effects is employed, such as Guass-Hermite quadrature rule.
In the M-step it is possible to obtain the estimation for σ_q^2 and D in closed form solution. For the others parameters there is not a close solution, so it is necessary to use one-step Newton-Raphson update:

$$\hat{\beta}^{it+1} = \hat{\beta}^{it} - \left\{ \frac{\partial}{\partial \beta} S(\hat{\beta}^{it}) \right\}^{-1} S(\hat{\beta}^{it}) \quad ; \quad \hat{\theta}_t^{it+1} = \hat{\theta}_t^{it} - \left\{ \frac{\partial}{\partial \theta_t} S(\hat{\theta}_t^{it}) \right\}^{-1} S(\hat{\theta}_t^{it})$$

where $\hat{\beta}^{it}$ and $\hat{\theta}_t^{it}$ denote the values of β and θ_t at the current iteration. In addition, $S(\hat{\beta}^{it})$ and $S(\hat{\theta}_t^{it})$ denote the corresponding blocks of the Hessian matrix, evaluated at $\hat{\beta}^{it}$ and $\hat{\theta}_t^{it}$, respectively. For the evaluation of the blocks of the Hessian matrix, the numerical derivative routine is used.
At convergence the standard errors are evaluate with the empirical information matrix [7]:

$$I_e(\theta) = \sum_{i=1}^{n} s_i s_i' - n^{-1} \left(\sum_{i=1}^{n} s_i \right) \left(\sum_{i=1}^{n} s_i \right)' \tag{3}$$

where $s_i = \frac{\partial l_i(\theta)}{\partial \theta}$.
We implemented the algorithm in R software. The outline of the algorithm follows the points:

1. The initial values are estimated through the two-stage approach.
2. In the E-step the expected value of the complete data log-likelihood function is used considering the random effects as the missing data using, in addition, Guass-Hermite quadrature rule
3. In the M-step, for σ_q^2 and D it is possible to obtain closed form solution, while for the parameter γ, α_q and β_q a one-step Newton-Raphson update is implemented. Random effects and the baseline hazard are updated.
4. Iterate between step 2 and 3 until the algorithm converges, when the parameter estimates become stable.
5. At convergence, the standard errors for each parameter are calculated using empirical information matrix.

3 Application to Primary Biliary Cirrhosis dataset

We apply the algorithm to the primary biliary cirrhosis dataset (PBCSEQ) that is available from the package *Survival* in R [9]. The dataset established from Mayo Clinic consists of 312 clinical trial patients with primary biliary cirrhosis [8] followed up from 1974 to 1986. For each patient multiple laboratory results were collected at each visit of the follow-up. After analysing several possible models, two longitudinal covariates are considered: the level of serum bilirubin in mg/dl (*serBilir*), and the level of albumin in mg/dl (*albumin*). The observational time is expressed in days. In the survival sub-model, the exogenous covariate patient's age at registration in years (*age*) is analysed. Accordingly the longitudinal and the survival sub-models used are:

$$\begin{cases} y_{i1}(t) = \beta_{01} + \beta_{11}t + b_{i01} + b_{i11}t + \varepsilon_{i1}(t) \\ y_{i2}(t) = \beta_{02} + \beta_{12}t + b_{i02} + b_{i12}t + \varepsilon_{i2}(t) \\ h_i(t) = h_0(t)\exp[\alpha_1 m_{i1}(t) + \alpha_2 m_{i2}(t) + \gamma_1 age] \end{cases}$$

where $y_{i1}(t)$ is the *log(serBilir)* and $y_{i2}(t)$ is the *albumin*.
The results obtained using the new algorithm implemented are shown in Table 1, where every parameter results to be statically significant.

Table 1 Results of the joint model on PBCSEQ dataset

Parameter	Est.	SE	p-value
α_1 (*log(serBilir)*)	1.1700	0.1052	< 0.0001
α_2 (*albumin*)	-1.8784	0.1557	< 0.0001
γ_1 (*age*)	0.0510	0.0075	< 0.0001
β_{01} (*Intercept*)	0.6371	0.0134	< 0.0001
β_{11} (*Time*)	0.0005	$8.3434{*}10^{-06}$	< 0.0001
β_{02} (*Intercept*)	3.5345	0.0201	< 0.0001
β_{12} (*Time*)	-0.0003	$1.1751{*}10^{-05}$	< 0.0001

Log-likelihood -2957.639

In particular, the *log(serBilir)* affects positively the risk of death (a one point increase in the *log(serBilir)* is associated with a 3.2220 $(= \exp(1.1700))$ fold increase in the risk of death), while the *albumin* affects negatively the risk of death (a one point increase the *albumin* will give a 0.1528 $(= \exp(-1.8784))$ fold decrease in the risk of death). Moreover the exogenous variable, *age*, affects positively the risk of death (one point increase in *age* gives a 1.0523 $(= \exp(0.0510))$ fold increase in the risk of death). Analysing the longitudinal sub-models, the observational time affects positively ($\beta_{11} = 0.0005$) the level of *log(serBilir)*, on the contrary it is negatively associated ($\beta_{12} = -0.0003$) with the level of *albumin*.

4 Conclusions and ideas of further work

The aim of the paper is to extend the maximum likelihood estimation method proposed by Rizopoulos [6] to the case in which the longitudinal sub-model is multivariate. We presented the algorithm and applied it to the PBCSEQ dataset.

The results are encouraging and deal to several ideas of future work. Developing, for instance, deeper diagnostic analysis and dynamic predictions. Another idea for further work is extending the survival sub-model, studying the joint effect of more than one longitudinal covariate on more than one terminal event.

References

1. P. Albert and J. Shih. An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics*, 4:1517–1532, 2010.
2. C. Faucett and D. Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, 15:1663–1685, 1996.
3. A. Henderson, V. De Gruttola, and M. Wulfsohn. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1:465–480, 2000.
4. G. Hickey, P. Philipson, A. Jorgensen, R. Kolamunnage-Dona, P. Williamson, and D. Rizopoulos. *joineRML: Joint Modelling of Multivariate Longitudinal Data and Time-to-Event Outcomes*, 2017. version 0.4.1.
5. G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16:117–131, 2016.
6. D. Rizopoulos. *Joint model for Longitudinal and Time-to-Event Data with applications in R*. CRC Press, Boca Raton, 2012.
7. A. Scott. Maximum likelihood estimation using the empirical fisher information matrix. *Journal of Statistical Computation and Simulation*, 72(8):599–611, 2002.
8. T. Therneau and P. Grambsch. *Modeling Survival Data: extending the Cox Model*. Springer-Verlang, New York, 2000.
9. T. Therneau and T. Lumley. *survival: Survival Analysis*, 2015. version 2.41-3.
10. J. Xu and S. Zeger. The evaluation of multiple surrogate endpoints. *Biometrics*, 57:81–87, 2001.
11. J. Xu and S. Zeger. Joint analysis of longitudinal data comprising repeated measures and times to events. *Applies Statistics*, 50:375–387, 2001.