Variance component estimation on Competing risk analysis with masked causes and gaussian random components: A simulation study.

Rafael Pimentel Maia*
ESALQ/USP, Piracicaba, Brazil - e-mail address: rafaelp.maia@usp.br

Clarice Garcia Demetrio Borges
ESALQ/USP, Piracicaba, Brazil - e-mail address: clarice.demetrio@usp.br

Rodrigo Labouriau
Aarhus University, Aarhus, Denmark - e-mail address: rodrigo.labouriau@mbg.au.dk

Abstract

The problem of competing risks arises in time-to-event analysis when the subjects observed may experience one among a set of possible competing events. For instance, in longevity studies often the interest lies in modelling the time until death of a group of individuals that might die of di erent speci c causes. The time to death of an individual is said to be *cause-masked*, or simply *masked*, when the time to death of this individual is observed but it is not known which of the possible causes of death occurred. This work will study some techniques, based on suitable variants of the EM-algorithm, to perform statistical inference in a competing risk scenario in the present day nce of partial masking and right censoring. We present an implementation of the EM-algorithm for treating the partial mixture induced by the masking of the causes of death. The goal is to extent a class of multivariate proportional hazard models for competing risks containing suitably de ned gaussian random components to characterize the quantitative genetic determination of longevity in large scale animal production systems. Moreover, it is provided a simulation study evaluating the performance of the proposed inferente procedures with respect the estimation of variance-covariance components. In this study we simulate a range of data of competing risks with three speci c causes of death based on a proportional hazard model with a binary xed e ect and a multivariate gaussian random component for each cause of death. Four scenarios were simulated representing di erent choices of the masked probabilities and the covariance structure of the random component. In genera, we concluded that the inference procedure based on the EM algorithm is able to detect part of variance of the random components but tended to under estimate the variances,I specially when the probability of masking were high.
Keywords: discrete time, cause-speci c hazard function, multivariate model, EM algorithm.

1. Introduction

The problem of competing risks arises in time-to-event analysis when the subjects observed may experience one among a set of possible competing events. For instance, in longevity studies often the interest lies in modelling the time until death of a group of individuals that might die of di erent speci c causes. The time to death of an individual is said to be *cause-masked*, or simply *masked*, when the time to death of this individual is observed but it is not known which of the possible causes of death occurred. This work will study some techniques, based on suitable variants of the EM-algorithm, to perform statistical inference in a competing risk scenario in the presence of partial masking and right censoring.

The problem of competing risks with masked causes has been treated in the literature by using a nite mixture sub-model to represent the masked individuals and then performing inference via the EM algorithm for nite mixtures (see Flehinger et al. 1996, 1998, 2002 and Craiu and Duchesne 2004a, 2004b). On the other hand, Maia et al. (2014a,b) used multivariate proportional hazard models for competing risks containing suitably de ned gaussian random components to characterize the quantitative genetic determination of longevity in large scale animal production systems. Here we will propose a methodology to combine these two techniques in order to characterize quantitative genetic aspects of traits involving right censure, competitive risks and partial masking. The speci c goals of this work are: a) to extend the class of complex models

proposed by Maia et al. (2014a) by treating the problem of partial masking via the EM algorithm, and b) to provide a simulation study evaluating the performance of the proposed models with respect the estimation of variance-covariance components.

The paper is organised in the following way. Section 2 presents a brief characterization of the problem and introduces the basic notation. The competing risk model proposed by Maia et. al. (2014) is described in Section 3. The conditional likelihood and the EM algorithm are described in Section 4. Section 5 presents the simulation algorithm and the results of the simulations for di erent scenarios are given in Section 6. A brief discussion and some conclusions are given in Section 7.

2. Notation and Problem Characterization

Assume that we observe n individuals where each of them may die of one, and only one, of J speci c causes of death. Here $J > 1$ and it is implicitly assumed that the set of those J causes of death is exhaustive, in the sense that no other cause of death may occur in the study in discussion. In the following we will index the n individuals by i, so we write X_i to represent the variables $X_1, \ldots, X_n$. Let T_i^* and C_i^* be two non-negative discrete random variables representing the survival time and the censure time of the i^{th} individual, respectively. Without loss of generality, we assume that T_i^* and C_i^* take values in $\mathbb{Z}_+$. Here the i^{th} individual is observed up to time $T_i = \min(T_i^*, C_i^*)$ and we say that an observation is *right-censored* when $T_i^* > C_i^*$. It is known whether the i^{th} observation is right-censored or not, so we observe the indicator variables $\delta_i = \mathbb{1}_{(T_i \leq C_i)}$. Moreover, when the time of death is known ($\delta_i = 1$) two things can happen: 1) the cause of death is known, or 2) it is not known which of the causes of death occurred (although it is known that the cause of death is one of the J causes); in the last case the cause of death is said to be *masked*. In order to keep track of the information on the deaths we introduce the variables D_i (for $i = 1, \ldots, n$) that takes the value j when the i^{th} individual dies by the cause j and 0 when the death cause is not known (whether because the i^{th} observation is right-censored or because the cause of death for the i^{th} individual is masked).

The *cause-specific hazard probability function* is de ned, for the i^{th} individual and the j^{th} cause, by

$$\lambda_{ij}(t) = P[T_i = t, D_i = j | T_i \geq t] \,. \tag{1}$$

Consider the random variable γ_i which indicates absence of masking, i.e., γ_i is equal to 1 when the true cause of death of the i^{th} individual is observed and 0 otherwise. With this notation, the probability that the cause of death of the i^{th} individual is masked is given by

$$\rho_j = P[\gamma_i = 0 | T_i, D_i = j, \delta_i = 1] \,. \tag{2}$$

Here we implicitly assume that the probabilities ρ_j do not change over the time. The probability that the true cause of death of the i^{th} is j given it has a masked cause is

$$\pi_{ij}(t) = P[D_i = j | T_i = t, \gamma_i = 0, \delta_i = 1] = \frac{\rho_j \lambda_{ij}(t)}{\sum_{k=1}^{J} \rho_k \lambda_{ik}(t)} \,. \tag{3}$$

3. The proportional hazard model

Suppose, additionally, that there is a range of explanatory variables (possibly time dependent) represented by the vectors $\mathbf{X}_i(t)$ and a gaussian random components, say $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_J)$ with a k dimensional component ($k \geq 1$) for each cause of death. The presence of the random component $\mathbf{U}$ in the models introduced below will enhance the exibility and applicability of those models. In particular, this will allow to adapt classical models of quantitative genetics to the context of censored models with competing risks (see Maia et al. 20014a, b) even in the presence of partial masking of the causes of death.

Based on the model proposed by Maia et. al. (2014a), the conditional cause speci c hazard function for the j^{th} cause ($j = 1, 2, \ldots, J$), conditional on $\mathbf{U}_j = \mathbf{u}_j$, for the i^{th} individual ($i = 1, 2, \ldots, n$) at the time t, $t \in \mathbb{Z}_+$, is given by

$$\lambda_{ij}(t | \mathbf{u}_j) = \lambda_j(t) \exp\left\{ \mathbf{X}_i^t(t)\boldsymbol{\beta}_j + \mathbf{Z}_i^t \mathbf{u}_j \right\} \,, \tag{4}$$

where the $\lambda_j(\cdot)$s are the baseline speci c hazard function and $\boldsymbol{\beta}_j$ are the vectors of xed e ects. $\mathbf{X}$(t) and $\mathbf{Z}$ are incidence matrices. It is assumed that $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_J)$ follows a multivariate normal distribution with mean equal to zero and covariance matrix given by $A \otimes \boldsymbol{\Sigma}$, where A is a known matrix (usually an identity matrix or a relationship matrix) and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1J} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1J} & \sigma_{2J} & \ldots & \sigma_J^2 \end{bmatrix}. \tag{5}$$

Denote the vector of parameters $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_J, \rho_1, \rho_2, \ldots, \rho_J)$ by $\boldsymbol{\theta}$. The marginal likelihood function for the respective model is given by

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \int \prod_{i=1} [1 - \lambda_{i\cdot}(t_i; \boldsymbol{\beta}|\mathbf{u})]^{1-\delta_i} S_i(t_i - 1; \boldsymbol{\beta}|\mathbf{u}) \prod_{j=1} \lambda_{ij}(t_i; \boldsymbol{\beta}_j|\mathbf{u}_j)^{\delta_{ij}} (1-\rho_j)^{\gamma_i \delta_{ij}} \rho_j^{(1-\gamma_i)\delta_{ij}} \phi(\mathbf{u}; \boldsymbol{\Sigma}) d\mathbf{u}\,, \tag{6}$$

where $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J\}$, t_i is the observed survived time for the i individual, $\lambda_{i\cdot}(t) = \sum_{j=1}^{J} \lambda_{ij}(t)$, $S_i(t) = \prod_{s<t} [1 - \lambda_{i\cdot}(s)]$ and $\phi(\cdot)$ is the multivariate normal probability density function. Note that the multiple integral above is typically of high dimension.

4. EM algorithm

We present here an implementation of the EM-algorithm for treating the partial mixture induced by the masking of the causes of death. First we set the initial values the parameters in the model, i.e. $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. The initial values for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are obtained by tting a multivariate proportional model where the individuals with masked causes are treated as right censored. The initial values for the parameters describing the masking probabilities $\rho_1, \ldots, \rho_J$, are given by $\rho_1^0 = \rho_2^0 = \ldots = \rho_J^0 = \frac{\#\text{ masked individuals}}{\#\text{ individuals not censured}}$. Other choices of initial values might easily be implemented.

In the E-step we calculate the expected value of δ_{ij} conditionally on the parameters estimates from the previous interaction of the algorithm, which given by

$$E\left[\delta_{ij}|\boldsymbol{\theta}^{l-1}, \boldsymbol{\Sigma}^{l-1}\right] = \begin{cases} 1 & \text{if } \gamma_i = 1 \text{ and } D_{ij} = j \\ 0 & \text{if } \gamma_i = 1 \text{ and } D_{ij} \neq j \text{ or if } \delta_i = 0 \\ \hat{\pi}_{ij}^l(t_i) & \text{if } \gamma_i = 0 \text{ and } \delta_i = 1\,, \end{cases} \tag{7}$$

where

$$\hat{\pi}_{ij}^l(t_i) = \frac{\hat{p}_j^{l-1} \lambda_{ij}(t_i; \hat{\boldsymbol{\beta}}_j^{l-1}|\mathbf{u}_j^{l-1})}{\sum_{k=1}^{J} \hat{p}_k^{l-1} \lambda_{ik}(t_i; \hat{\boldsymbol{\beta}}_k^{l-1}|\mathbf{u}_k^{l-1})}\,. \tag{8}$$

In the M-step we estimate $\boldsymbol{\theta}^l$ and $\boldsymbol{\Sigma}^l$ the values that maximise

$$Q(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\boldsymbol{\theta}^{l-1}, \boldsymbol{\Sigma}^{l-1}) = E\left[\log \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Sigma})|\boldsymbol{\theta}^l, \boldsymbol{\Sigma}^l\right]. \tag{9}$$

Note that $Q(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\boldsymbol{\theta}^{l-1}, \boldsymbol{\Sigma}^{l-1})$ is a function of $E[\delta_{ij}|\boldsymbol{\theta}^l, \boldsymbol{\Sigma}]$. The maximization above is performed based on a Laplace approximation for (6) and the use a proper multivariate generalized linear mixed model (see Breslow and Clayton (1993), and Maia et. al. (2014a) for more details). The algorithm stops when $\left\|\hat{\boldsymbol{\theta}}^l - \hat{\boldsymbol{\theta}}^{l-1}\right\| \leq \epsilon$, where $\|\cdot\|$ is the euclidean norm and ϵ is a small pre-selected value.

5. Simulation Study

In this study a range of data of competing risks with three speci c risks based on a proportional hazard model with a binary xed e ect and a multivariate gaussian random component will be simulated. Four scenarios were simulated representing di erent choices about the masked probabilities and the variance-covariance structure of the random component. The general characteristics common to the four scenarios are:

- The models involve three di erent causes, there are no censored observations and the baseline speci c hazard probability function was assumed to be constant;
- The cause speci c hazard probability, for the j$^{\text{th}}$ risk, is modelled by $\lambda_{ij}(t|\mathbf{u}) = \lambda_j \exp(\boldsymbol{\beta}_j X_i + Z_i \mathbf{u}_j)$
- The set values for the xed parameter are λ_1 = 0.10, λ_2 = 0.11, λ_3 = 0.12, $\boldsymbol{\beta}_1 = -0.12$, $\boldsymbol{\beta}_2 = -020$ and $\boldsymbol{\beta}_3 = -015$
- The sample sizes are $10,000$ individuals.

The simulations were performed using the software R version 3.1.2 and the M-step of the EM algorithm was calculated using the software DMU (see Madsen et. al. 2010 and Madsen et al 2014).

6. Results

Here we present the general results obtained for each of the four simulated scenarios. The rst three scenarios di er in terms of the complexity of the simulated models.

Scenario I

In the rst scenario we simulated datasets assuming equal masking probabilities, and four di erent values for ρ: 0.1, 0.25, 0.4 and 0.6). Moreover, we the random components associated to the three causes to be uncorrelated and the respective variances to be 0.12, 0.22 and 0.07, so that = $diag(0.12, 0.22, 0.07)$. Tables 1 and 2 display the estimates obtained ofr the masking probabilities and the variance components, respectively.

Table 1: Median and 95% con dence interval for the masking probabilities estimates at the Scenario I.

True ρ	**Median**	**95% IC**	
0.10	0.095	0.089	0.101
0.25	0.237	0.227	0.248
0.40	0.380	0.365	0.397
0.60	0.571	0.551	0.594

Table 2: Median and 95% con dence interval for the variance components estimates at the Scenario I.

		ρ = 0.1			ρ = 0.25		
Parameter	**True Value**	**Median**	**95% IC**		**Median**	**95% IC**	
σ_1^2	0.12	0.111	0.074	0.160	0.104	0.068	0.156
σ_2^2	0.22	0.208	0.143	0.294	0.200	0.139	0.279
σ_3^2	0.07	0.063	0.039	0.090	0.058	0.033	0.089
		ρ = 0.40			ρ = 0.60		
Parameter		**Median**	**95% IC**		**Median**	**95% IC**	
σ_1^2	0.12	0.097	0.058	0.147	0.077	0.029	0.130
σ_2^2	0.22	0.195	0.132	0.265	0.177	0.112	0.257
σ_3^2	0.07	0.052	0.026	0.083	0.033	0.000	0.071

Scenario II

The only difference from the scenario I to the scenario II was with respect the variance matrix for the random components. In the scenario II the random components was assumed to be correlated with covariance matrix given by $= \begin{pmatrix} 0.12 & -0.11 & 0.07 \\ -0.11 & 0.22 & -0.09 \\ 0.07 & -0.09 & 0.15 \end{pmatrix}$. Tables 3 and 4 present the estimates for the masking probabilities and the variance components, respectively. The models for ρ = 0.60 did not converged, probably because we did not have enough information to estimate the parameters for a model of such complexity.

Table 3: Median and 95% confidence interval for masking probabilities estimates at the Scenario II.

ρ	**Median**	**95% IC**	
0.10	0.095	0.088	0.101
0.25	0.237	0.227	0.247
0.40	0.380	0.365	0.393

Table 4: Median and 95% confidence interval for the variance components estimates at the Scenario II.

		ρ = 0.1			ρ = 0.25			ρ = 0.40		
Parameter	**True value**	**Median**	**95% IC**		**Median**	**95% IC**		**Median**	**95% IC**	
σ_1^2	0.12	0.111	0.074	0.158	0.106	0.067	0.155	0.100	0.056	0.151
σ_{21}	-0.11	-0.109	-0.161	-0.068	-0.108	-0.155	-0.065	-0.103	-0.160	-0.059
σ_2^2	0.22	0.208	0.146	0.287	0.202	0.139	0.282	0.200	0.129	0.282
σ_{31}	0.07	0.067	0.035	0.103	0.070	0.035	0.107	0.072	0.039	0.113
σ_{32}	-0.09	-0.088	-0.138	-0.045	-0.085	-0.138	-0.045	-0.083	-0.137	-0.035
σ_3^2	0.15	0.139	0.098	0.194	0.136	0.088	0.189	0.127	0.078	0.186

Scenario III

In the third scenario we assumed the same variance matrix considered in the scenario II but we allowing the masking probabilities to change according to the true cause of death. The values used were ρ_1 = 0.1, ρ_2 = 0.2, ρ_3 = 0.3. The estimates for the masking probabilities and the variance components are presented in the Tables 5 and 6, respectively.

Table 5: Median and 95% confidence interval for masking probabilities estimates at the Scenario III.

Parameter	**True Value**	**Median**	**95% IC**	
ρ_1	0.05	0.081	0.036	0.119
ρ_2	0.12	0.120	0.077	0.161
ρ_3	0.17	0.138	0.110	0.201

Table 6: Median and 95% con dence interval for the variance components estimates at the Scenario III.

Parameter	**True value**	**Median**	**95% IC**	
σ_1^2	0.12	0.110	0.072	0.158
σ_2^2	0.22	0.207	0.142	0.295
σ_3^2	0.07	0.063	0.039	0.094

Scenario IV

In the last scenario the dataset were generated assuming a constant masking probability equal to 0.20 and the model was tted assuming the masking probabilities to depend on the true cause of death. The results as displayed at the Tables 7 and 8.

Table 7: Median and 95% con dence interval for masking probabilities estimates at the Scenario IV.

Parameter	True value	Median	**95% IC**	
ρ_1	0.20	0.200	0.147	0.256
ρ_2	0.20	0.201	0.161	0.246
ρ_3	0.20	0.199	0.138	0.238

Table 8: Median and 95% con dence interval for the variance components estimates at the Scenario IV.

Parameter	True value	Median	**95% IC**	
σ_1^2	0.12	0.107	0.071	0.155
σ_2^2	0.22	0.203	0.140	0.292
σ_3^2	0.07	0.060	0.035	0.091

6. Conclusions

In general we conclude that the nite mixture model approach via EM algorithm is able to detect part of variance of the random components but tended to under estimate the variances specially when the probability of masking were high. Moreover, as we see at the results of the scenario II (where the complexity of the variance-covariance components matrix is larger) the estimation procedure is unstable for dataset with large proportion of masked individual, since there would be not enough information for estimate all the variance-covariance parameters in the model. We could also observe from the results of the last scenario that the model was able to estimate properly the masking probabilities even under the hypothesis of not constant masking probabilities when the dataset was generate based on a constant masking probabilities model.

References

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

Craiu, R. V. & Duchesne, T. (2004). Inference based on the EM algorithm for the competing risks model masked causes of failure. Biometrika, 91, pp. 543{558.

Craiu, R, V. & Reiser, B. (2006). Inference for the dependent competing risks model with masked causes of failure. Lifetime Data Analysis, 12, pp. 21{53.

Flehinger, B. J., Reiser, B. & Yashchin, E. (1996). Inference about defects in the presence of masking. Technometries, 38, pp. 247{255.

Flehinger, B. J., Reiser, B. & Yashchin, E. (1998). Survival with competing risks and masked causes of failures. Biometrika, 85, pp. 151{164.

Flehinger, B. J., Reiser, B. & Yashchin, E. (2002). Parametric modeling for survival with competing risks and masked failure causes. Lifetime Data Analysis, 8, pp. 177{203.

Madsen, P., Su, G., Labouriau, R. & Christensen, O. F. (2010). DMU A package for analyzing multivariate mixed models. Page 732 in Proc. 9th World Congress on Genetics Applied to Livestock Production (WCGALP), Leipzig, Germany. Gesellschaft fr Tierzuchwissenschaften e. V., Neustadt, Germany.

Maia, R. P., Madsen, P. & Labouriau, R. (2014a). Multivariate survival mixed models for genetic analysis of longevity trait. Journal of Applied Statistics. 42, pp. 1286{1306.

Maia, R. P., Ask, B., Madsen, P., Pedersen, J. & Labouriau, R. (2014b). Genetic determination of mortality rate in Danish dairy cows: A multivariate competing risks analysis based on the number of survived lactations. Journal of Dairy Science. 97, pp. 1753{1761.

Fisher, R. A. (1925). Statistical methods for research workers. Genesis Publishing Pvt Ltd.

McCullagh, P., & Nelder, J. A. (1983). Generalized linear models. London, England, Chapman and Hall.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2013). Time series analysis: forecasting and control. John Wiley & Sons.