A Regression Model for
the Copula Graphic Estimator

S.M.S. Lo and R.A. Wilke

April 2011

A Regression Model for the Copula Graphic Estimator.[*]

Simon M.S. Lo[†]

Ralf A. Wilke[‡]

March 2011

Abstract

We consider a dependent competing risks model with many risks and many covariates. We show identifiability of the marginal distributions of latent variables for a given dependence structure. Instead of directly estimating these distributions, we suggest a plug-in regression framework for the Copula-Graphic estimator which utilises a consistent estimator for the cumulative incidence curves. Our model is an attractive empirical approach as it does not require knowledge of the marginal distributions which are typically unknown in applications. We illustrate the applicability of our approach with the help of a parametric unemployment duration model with an unknown dependence structure. We construct identification bounds for the marginal distributions and partial effects in response to covariate changes. The bounds for the partial effects are surprisingly tight and often reveal the direction of the covariate effect.

Keywords: Archimedean copula, dependent censoring

1 Motivation

The estimation of the marginal distribution, F_j, of a latent competing random variable T_j, is of prime interest to determine causal relationships between a covariate and the time to an event. F_j can be estimated directly by maximum likelihood techniques if it is known up to some unknown coefficients and if the copula is known. Chen (2010) suggests a semiparametric transformation model which includes the proportional odds and proportional hazard model as special cases. As a drawback of this approach, direct estimation of F_j requires full or at least some knowledge about F_j. F_j can be also determined with the Copula-Graphic estimator (Zheng and Klein, 1995). It

exploits a 1-1 relationship between the cumulative incidence curve for T_j, Q_j, and F_j if the copula is known. By using a nonparametric estimator for Q_j it is possible to determine F_j, without having any prior knowledge about it. As an important limitation, this approach was not yet extended to a model with many covariates. Braekers and Veraverbeke (2005) consider a nonparametric model with one continuous covariate but to our knowledge a general regression framework for multiple covariates has still to be developed.

This paper closes this gap by suggesting a multiple regression model for the Copula-Graphic estimator. We are able to establish a direct link between a multivariate Q_j and F_j conditional to many covariates. Our approach works with any consistent estimator for Q_j and is therefore not restricted to specific subdistribution models. Special cases of our model therefore include semiparametric (Fine and Gray, 1999) or parametric models (Jeong and Fine, 2007) for Q_j. We elaborate in detail a parametric regression model for which we derive a closed form solution for F_j and its asymptotic covariance matrix. In particular, we consider a parametric maximum likelihood estimator for Q_j. This model includes the proportional odds and proportional hazard model with Gompertz baseline function as special cases (Jeong and Fine, 2007). We claim that our approach is appealing for empirical research as it does not impose direct restrictions on F_j which would be difficult to test. Instead, it impose restrictions on Q_j which can be directly verified by data. Our implementation of the estimation approach is fast as it is based on closed form solutions. STATA code is available on request from the first author. We demonstrate the applicability by means of a parametric unemployment duration model. As the dependence structure between risks is unknown in this example, we construct bounds for $S_j = 1 - F_j$ which are due to the non-identifiability of the competing risks model. Moreover, we construct bounds for partial effects on S_j in response to covariate changes. We find that these bounds are rather tight in our example and estimation

results are often informative about the sign of a covariate effect.

The next section considers the general model. In Section 3 we derive closed form expressions for the parametric model. Section 4 presents an application to unemployment duration data and the last section provides a final discussion.

2 The Model

We consider a model with $j = 1, \ldots, J$ competing random variables $T_j \in \mathbb{R}_+$ with an unknown marginal distribution function $F_j(t; \mathbf{x}) = pr(T_j \leq t; \mathbf{x}) \in [0, 1]$ and marginal survival function $S_j(t; \mathbf{x}) = 1 - F_j(t; \mathbf{x})$. $\mathbf{x} \in \mathbb{R}^K$ is a $K \times 1$ vector of observable covariates. Due to the competing risks structure it is only possible to observe $(T, \delta, \mathbf{x})$ with $\delta = \text{argmin}_j\{T_j\}$ and $T = min_j\{T_j\}$. Let $(T_i, \delta_i, \mathbf{x}_i)$ be $i = 1, \ldots, N$ realisations of $(T, \delta, \mathbf{x})$ and $Q_j(t) = pr(T_j \leq t, \delta = j; \mathbf{x})$ be the cumulative incidence curve for risk $j = 1, \ldots, J$. The cause specific hazard is $h_j(t; \mathbf{x}) = \lim_{\Delta t \to 0}(1/\Delta t)pr\{t \leq T \leq t + \Delta t, \delta = j | T \geq t, \mathbf{x}\} = Q'_j(t; \mathbf{x})/S(t; \mathbf{x})$ with $Q'(t; \mathbf{x}) = dQ_j(t; \mathbf{x})/dt$. $S(t; \mathbf{x}) = pr(T \geq t; \mathbf{x})$ is the survival function of the minimum.

Zheng and Klein (1995) show that S_j can be identified if the copula, $C(S_1, \ldots, S_J)$ with coefficients ω, is known. Their approach, known as the Copula-Graphic estimator, is restricted to a model with $J = 2$ and without $\mathbf{x}$. Lo and Wilke (2010) generalize the Copula-Graphic estimator to $J > 2$ if the copula is Archimedean. Given $Q_j(t)$ for all j and $S(t)$, $S_j(t)$ can be determined by solving a system of equations. In this paper we suggest a regression setting for the copula graphic estimator which does not restrict the number of risks nor the number of covariates. We therefore develop a regression setting for $S_j(t; \mathbf{x})$ given $Q_j(t; \mathbf{x})$ and $C(.)$.

In the regression setting, $S_j(t; \mathbf{x})$ can be identified using two approaches: First, it is possible to specify the joint likelihood function if $C(.)$ is known, and $S_j(t; \mathbf{x})$ belongs to a known parametric

or semiparametric family with unknown coefficients ψ_j for all j. For the purpose of illustration, we sketch the likelihood function for a two risks model:

$$
\begin{aligned}
L(\psi_1, \psi_2) &= \prod_{i=1}^{N} pr(T_{1i} = t, T_{2i} > t, \mathbf{x}_i)^{\mathbf{1}_{\{\delta_i=1\}}} pr(T_{1i} > t, T_{2i} = t, \mathbf{x}_i)^{\mathbf{1}_{\{\delta_i=2\}}} \\
&= \prod_{i=1}^{N} \prod_{j=1}^{2} \frac{\partial}{\partial S_j} C\{S_1(T_i; \mathbf{x}_i; \psi_1), S_2(T_i; \mathbf{x}_i; \psi_2)\}^{\mathbf{1}_{\{\delta_i=j\}}}.
\end{aligned}
$$

For fully specified $S_1(t; \mathbf{x}; \psi_1)$ and $S_2(t; \mathbf{x}; \psi_2)$, standard methods can be applied to estimate the unknown coefficients by maximising $L(.)$. Chen (2010) considers the case when $S_1(t; \mathbf{x}; \psi_1)$ and $S_2(t; \mathbf{x}; \psi_2)$ belong to a semiparametric transformation model which includes the proportional hazard and the proportional odds models as special cases. Nonparametric maximum likelihood estimators are applied to solve for the unknown coefficients. Note that direct specification of the joint distribution is not an extension of the copula graphic estimator. The main idea of the copula graphic estimator is that the marginal survival function can be identified without imposing any direct parametric or semiparametric structure on them.

The second approach is to generalise the Copula-Graphic estimator to a model with covariates. We are only aware of one such attempt. Braekers and Veraverbeke (2005) use nonparametric kernel estimators for $F_j(t; \mathbf{x})$ but their approach is limited to $K = 1$. In this paper we extend the Copula-Graphic estimator to a regression model with many covariates.

Identifiability We require two regularity conditions for the copula to show identifiability of the unknown $S_j(t; \mathbf{x})$ for known $Q_j(t, ; \mathbf{x})$ and known or assumed $C(S_1, \ldots, S_J)$.

Assumption 1 $C(S_1, \ldots, S_J)$ *has the following properties:*

 1. *It is Archimedean;*

 2. $C(S_1, \ldots, S_J) = C(S_1, \ldots, S_J, \mathbf{x})$.

Then

$$C(s_1, \ldots, s_J) = \phi^{-1}[\phi\{S_1(t;\mathbf{x})\}, \ldots, \phi\{S_J(t;\mathbf{x})\}]$$

with $\phi(S) : [0,1] \to \mathbb{R}_+$ is the so called copula generator. ϕ is a strictly decreasing and twice differentiable continuous function with $\phi(1) = 0$.

Proposition 1 $S_j(t;\mathbf{x})$ *is identified under Assumption 1 if* $Q_j(t;\mathbf{x})$ *is known for all j. There exists a closed form solution*

$$
\begin{aligned}
S_j(t;\mathbf{x}) &= \phi^{-1}\left[-\int_0^t \phi'\{S(u;\mathbf{x})\}S(u;\mathbf{x})h_j(u;\mathbf{x})du\right] \qquad (1)\\
&= \phi^{-1}\left[-\int_0^t \phi'(1 - \sum_{j=1}^J Q_j(u;\mathbf{x}))Q_j'(u;\mathbf{x})du\right]\\
&= \Psi_j(\mathbf{Q}(t;\mathbf{x})).
\end{aligned}
$$

with $\mathbf{Q}(t;\mathbf{x}) = (Q_1(t;\mathbf{x}), \ldots, Q_J(t;\mathbf{x}))'$.

See Appendix 1 for the proof. This result is a generalization of equation (7) in Rivest and Wells (2001) for a two risks model in absence of covariates. Q_j can be directly estimated from data without specifying the copula function. It therefore differs from the common approach to identification of the competing risks models which directly specifies and estimates $S_j(t;\mathbf{x})$. We claim that it is easier to check the specification of $Q_j(t;\mathbf{x})$ as it describes an observed rather than a latent quantity. It is therefore easier to verify consistent estimation of $Q_j(t;\mathbf{x})$ in an application than verifying the consistency conditions for direct estimation of $S_j(t;\mathbf{x})$. Our approach reverses the approach of Cheng, Fine and Wei (1998) who first estimate a model for $S_j(t;\mathbf{x})$ and then predict $Q_j(t;\mathbf{x})$.

Estimation and large sample properties Equation (1) forms the basis of our model for estimating $S_j(t;\mathbf{x})$. Since $C(.)$ is known, the right hand side is nothing else than a known function

of $Q_j(t;\mathbf{x})$ as $S(t;\mathbf{x}) = 1 - \sum_{j=1}^{J} Q_j(t;\mathbf{x})$ and $h_j(t;\mathbf{x}) = Q_j'(t;\mathbf{x})/S(t;\mathbf{x})$. It is therefore straightforward to estimate $S_j(t;\mathbf{x})$ by $\hat{S}_j(t;\mathbf{x})$ in a second stage after $Q_j(t;\mathbf{x})$ was first estimated by $\hat{Q}_j(t;\mathbf{x})$:

$$\hat{S}_j(t;\mathbf{x}) = \phi^{-1}\left[-\int_0^t \phi'\{\hat{S}(u;\mathbf{x})\}\hat{S}(u;\mathbf{x})\hat{h}_j(u;\mathbf{x})du \right], \tag{2}$$

where $\hat{S}(t;\mathbf{x}) = 1 - \sum_{j=1}^{J} \hat{Q}_j(t;\mathbf{x})$ and $\hat{h}_j(t;\mathbf{x}) = \hat{Q}_j'(t;\mathbf{x})/\hat{S}(t;\mathbf{x})$. The model is very general as $\hat{Q}_j(t;\mathbf{x})$ can be nonparametric, semiparametric or parametric. In any case the large sample properties of $\hat{Q}_j(t;\mathbf{x})$ determine the large sample properties of $\hat{S}_j(t;\mathbf{x})$ something that is elaborated in more detail below. If $Q_j(t;\mathbf{x})$ is parametric, there may be an analytical solution to equation (1) (see also next section) otherwise it is be obtained my means of numerical methods. The same is true for the marginal effect of say x_k on $S_j(t;\mathbf{x})$ which is $\partial S_j(t;\mathbf{x})/\partial x$. The marginal effect is often of prime interest in applications.

Proposition 2 *Suppose $\hat{Q}_j(t;\mathbf{x})$ converges in probability to $Q_j(t;\mathbf{x})$ for all j and $\hat{\mathbf{Q}}(t;\mathbf{x})$ converges to a distribution with $J \times J$ covariance matrix $\mathbf{\Sigma}_Q$. Then under Assumption 1, we have:*

1. *$\hat{S}_j(t;\mathbf{x})$ converges in probability to $S_j(t;\mathbf{x})$, and*

2. *$\hat{S}_j(t;\mathbf{x})$ has asymptotic variance*

$$var(S_j) = [\mathbf{\Psi}_j'(\mathbf{Q}(t;\mathbf{x}))]'\mathbf{\Sigma}_Q[\mathbf{\Psi}_j'(\mathbf{Q}(t;\mathbf{x}))]$$

with $\mathbf{\Psi}_j'(\mathbf{Q}(t;\mathbf{x})) = \partial\Psi_j(\mathbf{Q}(t;\mathbf{x}))/\partial\mathbf{Q}(t;\mathbf{x})$ is $J \times 1$.

The consistency result is an immediate consequence of the Continuous Mapping Theorem (Van der Vaart, 1998, Theorem, 2.3) provided that Ψ is continuous due to Assumption 1. The distribution can be obtained by a direct application of the Delta Method (Van der Vaart, 1998, Theorem, 3.1) provided that Ψ is differentiable due to Assumption 1.

3 Parametric Example

Let us now assume that $Q_j(t) = pr(T_j \leq t, \delta = j; \mathbf{x})$ for $j = 1, \ldots, J$ is known up to some coefficients $\mathbf{\Theta}_j \in R^L$ with $L \geq K$. Recent work for such a model include Jeong and Fine (2007) and Peng and Fine (2009) or Lambert (2007). As an example we use a direct parametric modeling approach for Q_j. In particular, we specify an odds rate transformation model with Gompertz distribution as the improper baseline subdistribution (Jeong and Fine, 2007). In this model

$$Q_j(t; \mathbf{x}) = 1 - [1 + \alpha_j \tau_j \exp(\mathbf{x}'\boldsymbol{\beta}_j)\{\exp(\rho_j t) - 1\}/\rho_j]^{-1/\alpha_j}. \tag{3}$$

is known up to $\mathbf{\Theta}_j = \{\boldsymbol{\beta}'_j, \rho_j, \alpha_j, l\tau_j\}'$ with $\rho_j \in \mathbb{R}$ and $\tau_j = \exp(l\tau_j) \in \mathbb{R}_+$ as the parameters of the Gompertz distribution and $\alpha_j \in \mathbb{R}$ and $\boldsymbol{\beta}_j \in \mathbb{R}^K$ as the parameters of the transformation model. The cause specific hazard of risk j is

$$h_j(t; \mathbf{x}) = \frac{B_j \exp(-A_j(1 + \alpha_j)/\alpha_j)}{\sum_{j=1}^J \exp(-A_j/\alpha_j) - J + 1}$$

with

$$A_j = \ln[1 + \alpha_j \tau_j \exp(\mathbf{x}'\boldsymbol{\beta}_j)\{\exp(\rho_j t) - 1\}/\rho_j];$$

$$B_j = \tau_j \exp(\mathbf{x}'\boldsymbol{\beta}_j) \exp(\rho_j t).$$

This model encompasses the proportional odds ($\alpha_j = 1$) and the proportional hazard model ($\alpha_j \to 0$) with Gompertz baseline as special cases. Even though by being rather general for a parametric model, it restricts the subdistribution hazard to be monotonic in t.

As the copula function we assume the Frank copula.

Remark 1 *The Frank copula has the following properties:*

$$\phi(t) = -\ln\left[\frac{\exp(-\omega t) - 1}{\exp(-\omega) - 1}\right] \quad ; \quad \phi'(t) = \frac{\omega \exp(-\omega t)}{\exp(-\omega t) - 1};$$

$$\phi^{-1}(u) = -\frac{1}{\omega}\ln[\exp(-u)\{\exp(-\omega) - 1\} + 1] \quad ; \quad \phi'^{-1}(u) = \frac{\exp(-u)\{\exp(-\omega) - 1\}}{\omega[\exp(-u)\{\exp(-\omega) - 1\} + 1]}$$

with $\omega \in (-\infty, \infty) \setminus \{0\}$.

When $\omega = 0$, we have $\phi(t) = -ln(t)$. In what follows we derive closed form expressions for $S_j(t; \mathbf{x})$ and the marginal effect $\partial S_j(t; \mathbf{x})/\partial x_k$ for given values of $\boldsymbol{\Theta}_j$ and for $\omega \notin 0$. For notational convenience we consider a two risks model as it is always possible to apply the risk pooling approach of Lo and Wilke (2010) if the copula is Archimedean. In this case the first risk is of direct interest and second is the minimum of all other risks which does not have a direct interpretation.

First note that

$$S(t; \mathbf{x}) = \exp(-A_1/\alpha_1) + \exp(-A_2/\alpha_2) - 1; \tag{4}$$

$$Q_j'(t; \mathbf{x}) = B_j \exp\{-A_j(1 + 1/\alpha_j)\}. \tag{5}$$

By plugging (4) and (5) into (1), we obtain a closed form solution for the latent marginal distribution:

$$S_j(t; \mathbf{x}) = \phi^{-1}[-I_j(t; \mathbf{x})], \tag{6}$$

where

$$I_j(t; \mathbf{x}) = \int_0^t D_j(u; \mathbf{x})du; \quad \text{and} \tag{7}$$

$$D_j(t; \mathbf{x}) = \phi'(S(t; \mathbf{x}))Q_j'(t; \mathbf{x}) = \frac{\omega \exp[-\omega S(t; \mathbf{x})]}{\exp[-\omega S(t; \mathbf{x})] - 1}Q_j'(t; \mathbf{x}). \tag{8}$$

Equation (6) suggests that imposing parametric structure on $Q_j(t; \mathbf{x})$ implies some shape restrictions for $S_j(t; \mathbf{x})$. As the functional form of S_j is non trivial, the interpretation of β_j is unclear. To study the effect of a change in $\mathbf{x}$ on the survival probability it is possible to compute $S_j(t; \mathbf{x})$ for different values of $\mathbf{x}$ and to compare the resulting probabilities. For a continuous covariate it

is also possible to consider the marginal effect which is

$$\partial S_j(t; \mathbf{x}) / \partial x_k \;\; = \;\; -\phi'^{-1}\{-I_j(t, \mathbf{x})\}\frac{\partial I_j(t, \mathbf{x})}{\partial x_k} \tag{9}$$

where

$$\partial I_j(t, \mathbf{x}) / \partial x_k \;\; = \;\; \int_0^t \frac{\partial D_j(u; \mathbf{x})}{\partial x_k} \, du, \tag{10}$$

and

$$\partial D_j(t; \mathbf{x}) / \partial x_k \;\; = \;\; Q'_j \frac{\partial \phi'\{S(t; \mathbf{x})\}}{\partial x_k} + \frac{\partial Q'_j}{\partial x_k} \phi'\{S(t; \mathbf{x})\};$$

which can be solved by replacing

$$\partial \phi'\{S(t; \mathbf{x})\} / \partial x_k \;\; = \;\; \frac{\omega^2 \exp(-\omega S(t; \mathbf{x}))}{\{\exp(-\omega S(t; \mathbf{x})) - 1\}^2} \frac{\partial S(t; \mathbf{x})}{\partial x_k};$$

$$\partial S(t; \mathbf{x}) / \partial x_k \;\; = \;\; -\sum_j \frac{\exp(-A_j/\alpha_j)}{\alpha_j} \{1 - \exp(-A_j)\} \beta_{jk};$$

$$\partial Q'_j(t; \mathbf{x}) / \partial x_k \;\; = \;\; [1 - (1 + 1/\alpha_j)(1 - \exp(-A_j))] Q'_j \beta_{jk}.$$

Note that $\partial A_j / \partial x_k = \{1 - \exp(-A_j)\}\beta_{jk}$ and $\partial B_j / \partial x_k = B_j \beta_{jk}$.

It is apparent that the marginal effect (9) depends on t and $\mathbf{x}$ in a non trivial way. In our practical work with the model we have seen cases where the sign of the marginal effect changes in t. The parametric model (3) for Q_j therefore does not imply a non crossing property for marginal survival curves as it is assumed in the proportional hazards model for S_j. In an application the marginal effect may be determined at the sample mean of the covariates or as the average population marginal effect (AME) which is $AME = N^{-1}\sum_{i=1}^{N} \partial S_j(t; \mathbf{x}_i) / \partial x_k$.

Estimation and large sample properties The unknown coefficients $\boldsymbol{\Theta}_j$ for $j = 1, 2$ in (3) can be directly estimated by parametric maximum likelihood (Jeong and Fine, 2007). In this case

$\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2)'$ is consistent and asymptotically normal with known asymptotic covariance matrix. The estimator for $S_j(t; \mathbf{x})$ is a plug in solution to (2).

As this model is a special case of the model considered in Section 2, we know that the estimator for $S_j(t; x)$ is also consistent and asymptotically normal. In the following we derive a closed form for its asymptotic variance. First note that the asymptotic variance of $\hat{Q}_j(t; \mathbf{x}) = Q_j(t; \mathbf{x}, \hat{\boldsymbol{\Theta}}_j)$ is (Jeong and Fine, 2007)

$$\boldsymbol{\Sigma}_{\mathbf{Q}_j} = \left[\frac{\partial \mathbf{Q}(t; \mathbf{x})}{\partial \boldsymbol{\Theta}'_j}\right] \boldsymbol{\Sigma}_{\boldsymbol{\Theta}_j} \left[\frac{\partial \mathbf{Q}(t; \mathbf{x})}{\partial \boldsymbol{\Theta}'_j}\right]', \tag{11}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}_j}$ as the covariance matrix of $\boldsymbol{\Theta}_j$. For simplicity let us consider a two risks model ($J = 2$) but this can be easily extended to $J > 2$ under Assumption 1. Then, let $\mathbf{S} = (S_1, S_2)'$ and $\partial \mathbf{S}(t; \mathbf{x})/\partial \mathbf{x} = (\partial S_1(t; \mathbf{x})/\partial \mathbf{x}, \partial S_2(t; \mathbf{x})/\partial \mathbf{x})'$ with the latter being a $2K \times 1$ matrix. As the marginal distributions and the marginal effects are nonlinear functions of $\mathbf{Q}(t; \mathbf{x})$ we can write

$$\hat{\mathbf{S}}(t; \mathbf{x}) = \boldsymbol{\Psi}(\hat{\mathbf{Q}}(t; \mathbf{x})) = \boldsymbol{\Psi}^*(t; \mathbf{x}, \hat{\boldsymbol{\Theta}}) \tag{12}$$

$$\partial \hat{\mathbf{S}}(t; \mathbf{x})/\partial \mathbf{x} = \boldsymbol{\Xi}(\hat{\mathbf{Q}}(t; \mathbf{x})) = \boldsymbol{\Xi}^*(t; \mathbf{x}, \hat{\boldsymbol{\Theta}}). \tag{13}$$

Due to the consistency of $\hat{\mathbf{S}}$ we can focus on its asymptotic distribution with $\hat{\Theta} = \Theta|_{\Theta_0}$ and Θ_0 being the true value of Θ. The asymptotic covariance matrices of $\hat{\mathbf{S}}(t; \mathbf{x})$ and $\partial \hat{\mathbf{S}}(t; \mathbf{x})/\partial \mathbf{x}$ are

$$\boldsymbol{\Sigma}_{\mathbf{S}(t;\mathbf{x})} = \left[\frac{\partial \boldsymbol{\Psi}^*(t; \mathbf{x}, \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}'}\right] \boldsymbol{\Sigma}_{\boldsymbol{\Theta}} \left[\frac{\partial \boldsymbol{\Psi}^*(t; \mathbf{x}, \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}'}\right]'; \tag{14}$$

$$\boldsymbol{\Sigma}_{\partial \mathbf{S}(t;\mathbf{x})/\partial \mathbf{x}} = \left[\frac{\partial \boldsymbol{\Xi}^*(t; \mathbf{x}, \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}'}\right] \boldsymbol{\Sigma}_{\boldsymbol{\Theta}} \left[\frac{\partial \boldsymbol{\Xi}^*(t; \mathbf{x}, \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}'}\right]', \tag{15}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ as the covariance matrix of $\boldsymbol{\Theta}$. We show in Appendix 2 that there are closed form solutions for all right hand side components in (14) and (15).

4 Application

As an example to illustrate the applicability of our suggested regression model we apply the parametric Gompertz model for Q_j (Jeong and Fine, 2007) to unemployment duration data. We have considered this model in detail in Section 3. We use the data of McCall (1996) for our analysis which is also the illustrating data for competing risks models in the textbook of Cameron and Trivedi (2005). These data from the U.S. Current Population Survey's Displaced Worker Supplement provides information on three destination states for the displaced jobless individual: 1) full-time employment, 2) part-time employment and 3) either full or part-time employment but information on the employment type is missing. We focus on risk 1 and pool risks 2 and 3. Since not all joblessness periods were terminated by the time of the interview there is also independent censoring. In total there are 3,343 observations with 1,073 transitions into full-time employment and 913 transitions into the pooled risk. The Gompertz model includes time decreasing, time constant (exponential distribution) and time increasing subdistribution hazards as special cases. In unemployment duration we expect that the subdistribution hazard decreases over time. In the case of a decreasing hazard rate the corresponding probability distribution is improper which makes it an ideal candidate for the cumulative incidence curves. In an unemployment duration analysis this happens if a proportion of the unemployed individuals do not experience the event of interest (exit from unemployment) at all. Still the model for Q_j is parametric and may therefore not be able to capture all aspects of the shape of the true Q_j. It is, however, apparent from Figure 1 that the parametric model for Q_j is rather similar to the nonparametric estimator.

We estimate the model with seven covariates: $female$, $married$, $child$, $nonwhite$, $schgt12$ (more than 12 years of schooling), $houshead$ and ui (claiming unemployment insurance). The results of the parametric maximum likelihood estimator for Q_j are given in Table 1.

Table 1: Results of Maximum Likelihood estimator for the Gompertz model. Risk: Full time employment. Results for the other risk are not presented.

| Variable	Coefficient	(Std. Err.)
female	0.210*	(0.104)
married	0.401**	(0.106)
child	-0.010	(0.099)
nonwhite	-0.691**	(0.149)
schgt12	0.357**	(0.094)
houshead	0.373**	(0.105)
ui	-1.076**	(0.113)
ρ_1	0.005	(0.015)
α_1	2.972**	(0.380)
$ln(\tau_1)$	-2.719**	(0.134)
N		3,343
Log-likelihood		-7,747.359
$\chi^2_{(7)}$		138.79

Legend: *significant at 5%, **significant at 1%

Figure 1: Cumulative incidence for full time employment. Estimation without covariates
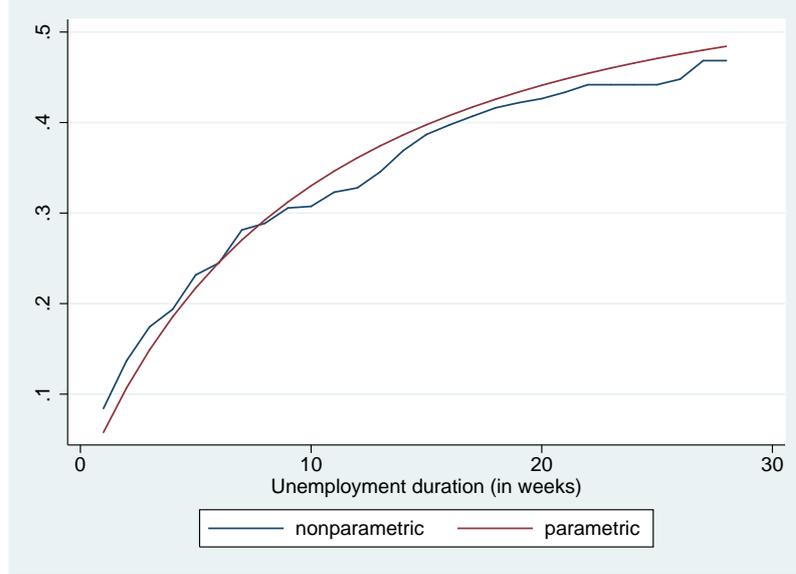
As a next step we explore whether $\hat{S}_j(t; \mathbf{x})$ varies with the choice of the dependence structure. As the copula we assume a Frank Copula but we allow the copula parameter ω to vary. Figure 2 shows the upper and the lower bound of the estimated marginal survival curve evaluated at $\mathbf{x} = \bar{\mathbf{x}}$. The lower bound is $min_\omega S_1(t; \mathbf{x}; \omega)$ and the upper bound is $max_\omega S_1(t; \mathbf{x}; \omega)$. We estimate $S_1(t; \mathbf{x}; \omega)$ for a fixed grid for ω. Instead of setting ω we use a grid for the Kendall tau rank correlation coefficient $\tau \in [-0.90, 0.90]$ and take the min and max of S_j over all grid points given t and $\mathbf{x}$. It is evident from Figure 2 that the bounds due to the non-identification of the competing risks model increase with elapsed duration and they amount to up to around 40 percentage points in this example. The estimated survival curve in the case of independent risks is located between the bounds but for shorter durations it is much closer to the upper bound. The 95% asymptotic confidence intervals for the identification bounds are quite close to the point estimate. The figure suggests that it is difficult to predict $S_j(t; \mathbf{x})$ precisely if the dependence structure is unknown. In particular the non-identification problem appear to be more problematic than the random sampling error. This observation is similar to the findings of Lo and Wilke (2010).

14

Figure 2: Bounds for $S_1(t; \bar{\mathbf{x}})$ for risk full time employment.
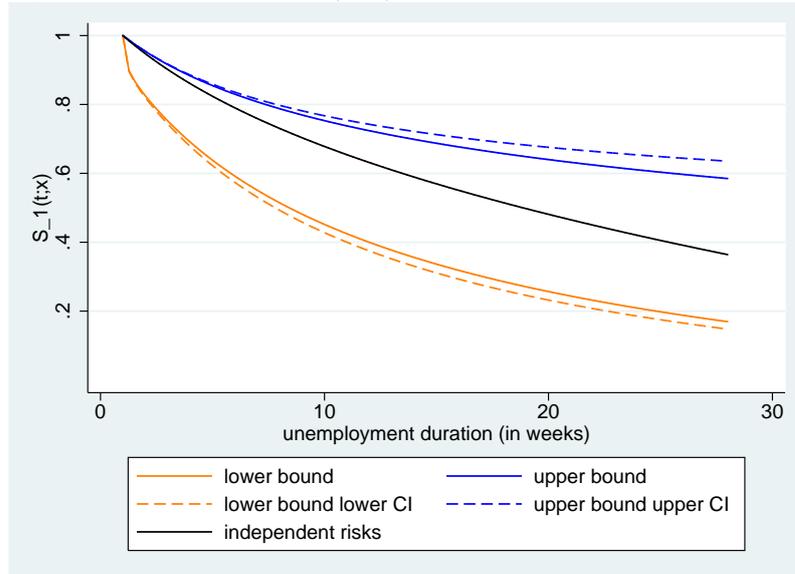
Lo and Wilke (2010) also find that changes in the marginal survival curves due to changes in covariates vary less with respect to the specification of the dependence structure and the sign of the change is mostly invariant. As a next step we therefore explore whether this is also true for our regression model. We analyse the partial effect due to changing a discrete variable from 0 to 1 with all other variables at their sample means. More specifically, we compute $S_1(t; \bar{\mathbf{x}}_{k1}; \omega) - S_1(t; \bar{\mathbf{x}}_{k0}; \omega)$ with $\bar{\mathbf{x}}_{km} = (\bar{x}_1, ..., \bar{x}_{k-1}, m, \bar{x}_{k+1}, ..., \bar{x}_7)'$. We then take the lower and upper bounds of these partial effects over all values of Kendall τ at each t and for all k. The resulting bounds for the partial effects for the seven variables are presented in Figure 3 along with their 95% lower and upper asymptotic confidence intervals. We make the following observations: First, the bounds for the partial effect tend to be tighter than the bounds for $S_1(t; \mathbf{x})$. Second, as the bounds do not contain zero in most cases, the sign of the partial effect is in all cases identified except for *child* for most durations. The latter finding is compatible with the results of Lo and Wilke (2011) who use stratified samples rather than a regression framework. While Lo and Wilke's (2011) analysis is restricted to the identifiability of the sign of the effect, our model suggests rather narrow bounds

for the magnitude. The results in Figure 3 also provide evidence that the estimated partial effect is often statistically significant. Third, although the bounds for the partial effect for *child* contain zero, we observe that the direction of the effect actually changes with elapsed duration for a given ω. As the position of the sign change varies with ω, the bounds in Figure 3(c) contain the value zero for most durations. A change in the direction of a covariate effect would not be obtained when directly fitting a proportional hazards model for the marginal distribution which would assume non-crossing of marginal survival curves.

Figure 3: Bounds for the partial effect of a regressor unit change on $S_1(t; \mathbf{x})$. Note: other regressors at their sample means.
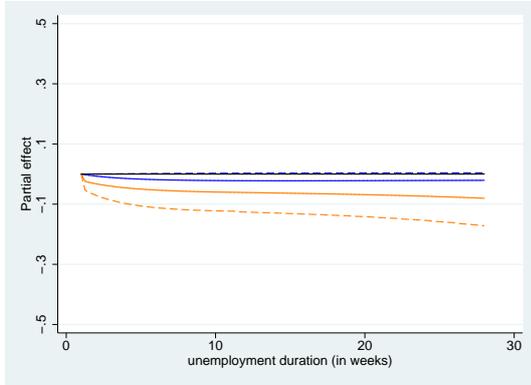
(a) *female*   (b) *married*
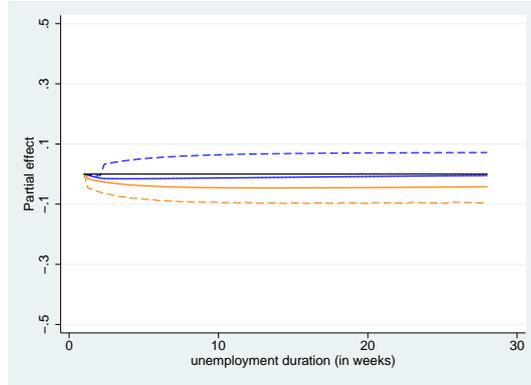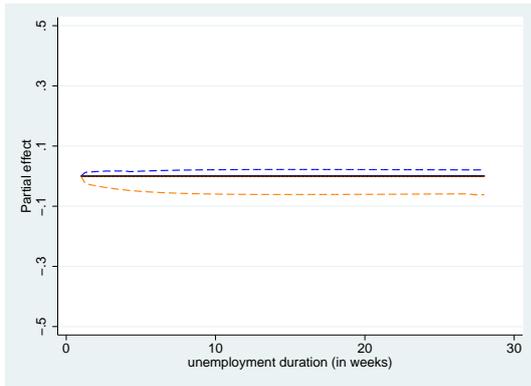
(c) *child*   (d) *nonwhite*
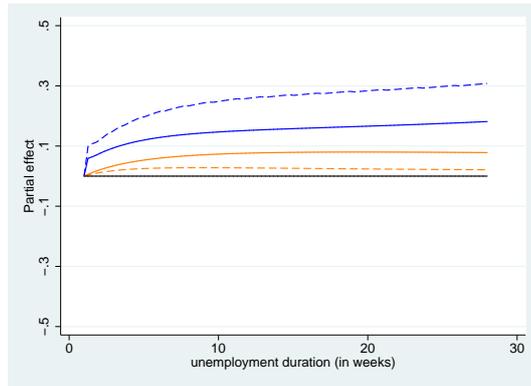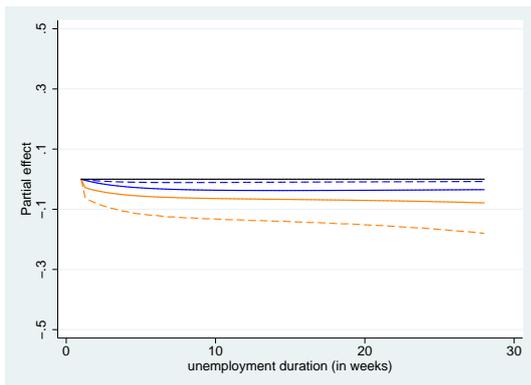
(e) *schgt*12   (f) *houshead*
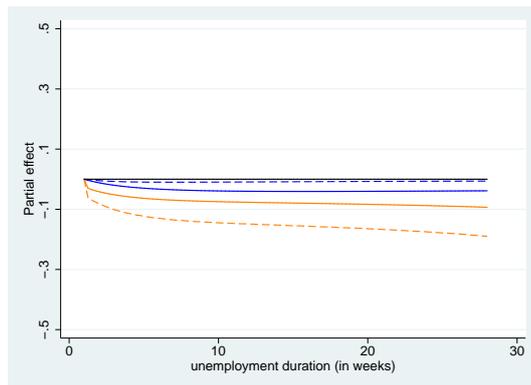
(g) *ui*
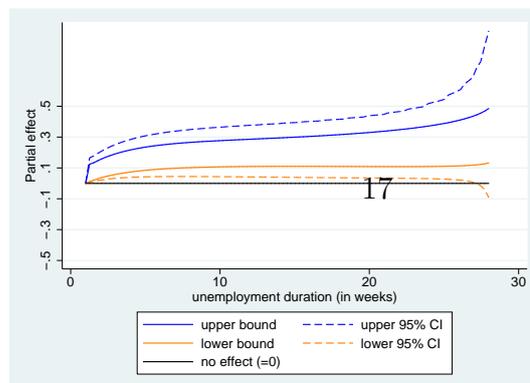
5 Discussion

We suggest a regression framework for the copula graphic estimator and therefore extend the model of Zheng and Klein (1995) and Lo and Wilke (2010) to many covariates. Under mild conditions we show identifiability and derive nice asymptotic properties for our framework. Our approach utilises direct estimates of the cumulative incidence curves which are identifiable quantities. It is therefore easier to verify the specification of the estimator than for estimating latent quantities. The only crucial assumption is that the model requires knowledge of the dependence structure between risks. In an application with unknown dependence structure, however, it is easy to create bounds for values of interest which represent the effect of the choice of the dependence structure. In our illustrative application of a parametric model to unemployment duration data we obtain that these bounds for partial effects are rather narrow while the bounds for the latent marginal distributions are wide. The results of our application therefore suggest that it is possible to obtain conclusive information about the effect of covariates even if the dependence structure is unknown.

References

[1] Braekers, R. and Veraverbeke, N. (2005) A copula-graphic estimator for the conditional survival function under dependent censoring, *Canadian Journal of Statistics*, 33, 429–447.

[2] Cameron, A.C., and Trivedi, P.K. (2005) *Microeconometrics*, Cambridge University Press, New York.

[3] Chen, Y.H. (2010) Semiparametric marginal regression analysis for dependent competing risks under an assumed copula, *Journal of the Royal Statistical Society B*, 72, 235–231.

[4] Cheng, S.C., Fine, J.P. and Wei, L.J (1998) Prediction of Cumulative Incidence Function under the Proportional Hazards Model, *Biometrics*, 54, 219–228.

[5] Fine, J.P. and Gray, R.J. (1999) A Proportional Hazards Model for the Subdistribution of a Competing Risk, *Journal of the American Statistical Association*, 94, 496–509.

[6] Jeong, J.H. and Fine, J.P. (2007) Parametric regression on cumulative incidence function, *Biostatistics*, 8, 1–13.

[7] Lambert, P.C. (2007) Modelling of the Cure Fraction in Survival Studies, *The Stata Journal*, 7, 351–375.

[8] Lo, S.M.S. and Wilke, R.A. (2010), A copula model for dependent competing risks, *Journal of the Royal Statistical Society Series C*, 59, 359–376.

[9] Lo, S.M.S. and Wilke, R.A. (2011), Identifiability and estimation of the sign of a covariate effect in the competing risks model, *mimeo*.

[10] McCall, B.P. (1996) Unemployment Insurance Rules, Joblessness, and Part-Time Work, *Econometrica*, 64, 647–682.

[11] Rivest, L. and Wells, M.T. (2001) A Martingale Approach to the Copula-Graphic Estimator for the Survival Function under Dependent Censoring, *Journal of Multivariate Analysis*, 79, 138–155.

[12] Van der Vaart, A.W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge.

[13] Zheng, M. and Klein, J.P. (1995), Estimates of marginal survival for dependent competing risks based on assumed copula. *Biometrika*, 82, 127–138.

Appendix 1

Proof of Proposition 1. For notational simplicity we present the case $J = 2$. The cause-specific hazard can be written as

$$
\begin{aligned}
h_1(t) = \frac{pr(T_1 = t, T_2 \geq t)}{pr(T_1 \geq t, T_2 \geq t)} &= \frac{-\frac{\partial}{\partial u} pr(T_1 \geq u, T_2 \geq v)|_{v=u}}{pr(T_1 \geq t, T_2 \geq t)} \\
&= -\frac{\partial S_1(t)}{\partial t} \frac{\partial C}{\partial S_1} \frac{1}{S(t)}
\end{aligned}
$$

When differentiating $\phi\{S(t)\} = \phi\{S_1(t)\} + \phi\{S_2(t)\}$ with respect to S_1 we obtain

$$
\phi'\{S(t)\}\frac{\partial C}{\partial S_1} = \phi'\{S_1(t)\}.
$$

We obtain (1) when combining the two above results:

$$
\begin{aligned}
\phi\{S_j(t)\} &= \int_0^t \phi'\{S(u)\}S(u)\frac{\partial S_1(u)}{\partial u}\frac{\phi'\{S_1(u)\}}{\phi'\{S(u)\}}\frac{1}{S(u)} \, du \\
&= \int_0^t \phi'\{S_1(u)\} \, dS_1(u).
\end{aligned}
$$

Appendix 2

Closed form of the asymptotic variance for the parametric model. We now show that there is a closed form solution of the asymptotic variance of the copula graphic regression estimator if Q_j is as in Equation (3) and estimated by Maximum Likelihood. First note that due to (7) and (8) we have in Equation (14)

$$\frac{\partial \Psi_j^*}{\partial \Theta'} = -\phi'^{-1}\{-I_j(t,\mathbf{x})\} \int_0^t \frac{\partial D_j(u;\mathbf{x})}{\partial \Theta'} \, du \quad \text{for } j = 1,2, \text{ with} \tag{16}$$

$$\frac{\partial D_j(u;\mathbf{x})}{\partial \Theta'} = Q_j' \frac{\partial \phi'\{S(t;\mathbf{x})\}}{\partial \Theta'} + \frac{\partial Q_j'}{\partial \Theta'} \phi'\{S(t;\mathbf{x})\}. \tag{17}$$

All components in the right hand side of equations (16) and (17) can be obtained analytically. They are given by:

$$\partial \phi'\{S(t;\mathbf{x})\}/\partial \Theta' = \frac{\omega^2 \exp(-\omega S(t;\mathbf{x}))}{\{\exp(-\omega S(t;\mathbf{x})) - 1\}^2} \frac{\partial S(t;\mathbf{x})}{\partial \Theta'};$$

$$\partial S(t;\mathbf{x})/\partial \Theta'_j = \exp\{-A_j/\alpha_j\}\mathbf{\Gamma}'_{1j};$$

$$\partial Q_j'(t;\mathbf{x})/\partial \Theta'_j = Q_j'\mathbf{\Gamma}'_{2j};$$

$$\partial Q_j'(t;\mathbf{x})/\partial \Theta'_i = \mathbf{0}, \forall i \neq j;$$

with $\mathbf{\Gamma}_{1j} =$

$$\begin{pmatrix} \{\exp(-A_j) - 1\}\mathbf{x}, \\ \{\exp(-A_j) - 1\}\{\rho_j t \exp(\rho_j t)\{\exp(\rho_j t) - 1\}^{-1} - 1\}/\{\alpha_j \rho_j\}, \\ \{A_j + \exp(-A_j) - 1\}/\alpha_j^2, \\ \{\exp(-A_j) - 1\}/(\alpha_j) \end{pmatrix};$$

and with $\boldsymbol{\Gamma}_{2j} =$

$$
\begin{pmatrix}
\{1 - (1 + 1/\alpha_j)(1 - \exp(-A_j))\}\mathbf{x}, \\
t - (1 + 1/\alpha_j)(1 - \exp(-A_j))\{\rho_j t \exp(\rho_j t)\{\exp(\rho_j t) - 1\}^{-1} - 1\}/\rho_j, \\
\{A_j + (1 + \alpha_j)(1 - \exp(-A_j))\}/\alpha_j^2, \\
\{1 - (1 + 1/\alpha_j)(1 - \exp(-A_j))\}
\end{pmatrix}.
$$

Note that

$$
\begin{aligned}
\partial A_j/\partial \boldsymbol{\beta}_j &= \{1 - \exp(-A_j)\}\mathbf{x}; \\
\partial A_j/\partial \rho_j &= \{1 - \exp(-A_j)\}[\rho_j t \exp(\rho_j t)/\{\exp(\rho_j t) - 1\} - 1]/\rho_j \\
\partial A_j/\partial \alpha_j &= \{1 - \exp(-A_j)\}/\alpha_j; \\
\partial A_j/\partial \tau_j &= \{1 - \exp(-A_j)\}; \\
\partial A_j/\partial \boldsymbol{\Theta}_i &= \mathbf{0}, \forall i \neq j; \\
\partial B_j/\partial \boldsymbol{\beta}_j &= B_j \mathbf{x}; \\
\partial B_j/\partial \rho_j &= B_j/\rho_j; \\
\partial B_j/\partial \alpha_j &= 0; \\
\partial B_j/\partial \tau_j &= B_j; \\
\partial B_j/\partial \boldsymbol{\Theta}_i &= \mathbf{0}, \forall i \neq j.
\end{aligned}
$$

Similarly we now show that there is an analytical solution for the right hand side of Equation (15). First, note that $\boldsymbol{\Xi}^*(t; \mathbf{x}, \hat{\boldsymbol{\Theta}}) = [\boldsymbol{\Xi}_1^{*\prime}, \boldsymbol{\Xi}_2^{*\prime}]'$. We use again (7) and (8) to obtain:

$$
\frac{\partial \Xi_j^*}{\partial \boldsymbol{\Theta}'} = -\phi'^{-1}\{-I_j\}\frac{\partial}{\partial \boldsymbol{\Theta}'}\left(\frac{\partial I_j}{\partial \mathbf{x}}\right) + \phi''^{-1}\{-I_j\}\frac{\partial I_j}{\partial \mathbf{x}}\frac{\partial I_j}{\partial \boldsymbol{\Theta}'},
$$

with

$$\frac{\partial}{\partial\mathbf{\Theta}'}\left(\frac{\partial I_j}{\partial\mathbf{x}}\right)$$
$$= \int_0^t \left[Q_j'\frac{\partial}{\partial\mathbf{\Theta}'}\left(\frac{\partial\phi'(S)}{\partial\mathbf{x}}\right) + \frac{\partial\phi'(S)}{\partial\mathbf{x}}\frac{\partial Q_j'}{\partial\mathbf{\Theta}'} + \frac{\partial}{\partial\mathbf{\Theta}'}\left(\frac{\partial Q_j'}{\partial\mathbf{x}}\right)\phi'(S) + \frac{\partial Q_j'}{\partial\mathbf{x}}\frac{\partial S}{\partial\mathbf{\Theta}'}\phi''(S) \right] du$$

and

$$\frac{\partial}{\partial\mathbf{\Theta}'}\left(\frac{\partial\phi'(S)}{\partial\mathbf{x}}\right) = \omega^2\exp(-\omega S)\{\exp(-\omega S)-1\}^2$$
$$\times\left[\frac{\partial}{\partial\mathbf{\Theta}'}\left(\frac{\partial S}{\partial\mathbf{x}}\right) + \omega\frac{\partial S}{\partial\mathbf{x}}\frac{\partial S}{\partial\mathbf{\Theta}'}\{2\exp(-2\omega S)-2\exp(-\omega S)-1\}\right]$$
$$\frac{\partial}{\partial\mathbf{\Theta}_j'}\left(\frac{\partial S}{\partial\mathbf{x}}\right) = \exp(-A_j/\alpha_j)\mathbf{\Gamma}_{3j}'/\alpha_j$$
$$\frac{\partial}{\partial\mathbf{\Theta}_j'}\left(\frac{\partial Q_j'}{\partial\mathbf{x}}\right) = \{1-(1+1/\alpha_j)(1-\exp(-A_j))\}Q_j'\boldsymbol{\beta}_j\mathbf{\Gamma}_{2j}'$$
$$-(1+1/\alpha_j)\exp(-A_j)Q_j'\boldsymbol{\beta}_j\frac{\partial A_j}{\partial\mathbf{\Theta}_j'}$$

and with $\mathbf{\Gamma}_{3j}=$

$$\begin{pmatrix} -\{1-\exp(-A_j)\}[\mathbb{1}+\{\exp(-A_j)+\{\exp(-A_j)-1\}/\alpha_j\}\mathbf{x}\boldsymbol{\beta}_j'], \\[6pt] (\partial A_j/\partial\rho_j)[\{1-\exp(-A_j)\}/\alpha_j-\exp(-A_j)]\boldsymbol{\beta}_j', \\[6pt] \{1-\exp(-A_j)\}[\{1-\exp(-A_j)-A_j\}/\alpha_j-\exp(-A_j)+1]\boldsymbol{\beta}_j'/\alpha_j, \\[6pt] -\{1-\exp(-A_j)\}[\exp(-A_j)-\{1-\exp(-A_j)\}/\alpha_j]\boldsymbol{\beta}_j' \end{pmatrix}.$$