Semiparametric Copula-Based Regression Models

Giampiero Marra
Statistical Science, University College London
joint work with Rosalba Radice, Birkbeck

11^{th} July 2017
38th Annual Conference of the International Society for Clinical Biostatistics

Intro

Copula-based regression models allow one to model jointly some outcomes of interest, where all the model's parameters can be modelled as flexible functional forms of covariate effects.

All models developed for the past 8-ish years have been incorporated in the `GJRM R` package, to facilitate the use of such models in industry and academia, and to enhance reproducible research.

First developments were motivated by the issues of *endogeneity* and *non-random sample selection* (or *unobserved confounding*). Examples: HIV prevalence estimation in the presence of missingness not at random, and estimating the impact of being a pro-active care patient on chronic depression in RCTs affected by partial compliance.

Intro - cont'd

Since then we have considerably extended the scope of the models implemented in the package.

In a nutshell, we can fit bivariate copula models under various sampling schemes (endogeneity, non-random sample selection, correlated errors' equations, partial observability), where the marginal responses can be binary, discrete, and continuous, and the effects of covariates can be specified using flexible additive predictors. We can also fit univariate models and started working on trivariate extensions and beyond.

Today, I will discuss bivariate **survival** models and illustrate them by estimating the dependence between colon cancer recurrence and death (such dependence is of particular interest since it may suggest the presence of certain underlying processes influencing the joint and conditional probabilities of the events).

The model

Consider right censored data where (U_{1i}, U_{2i}) represents censoring times which are independent of the survival times (T_{1i}, T_{2i}). We observe $(Y_{1i}, Y_{2i}) = (\min\{T_{1i}, U_{1i}\}, \min\{T_{2i}, U_{2i}\})$ and the censoring indicators $(u_{1i}, u_{2i}) = (I\{T_{1i} \leq U_{1i}\}, I\{T_{2i} \leq U_{2i}\})$.

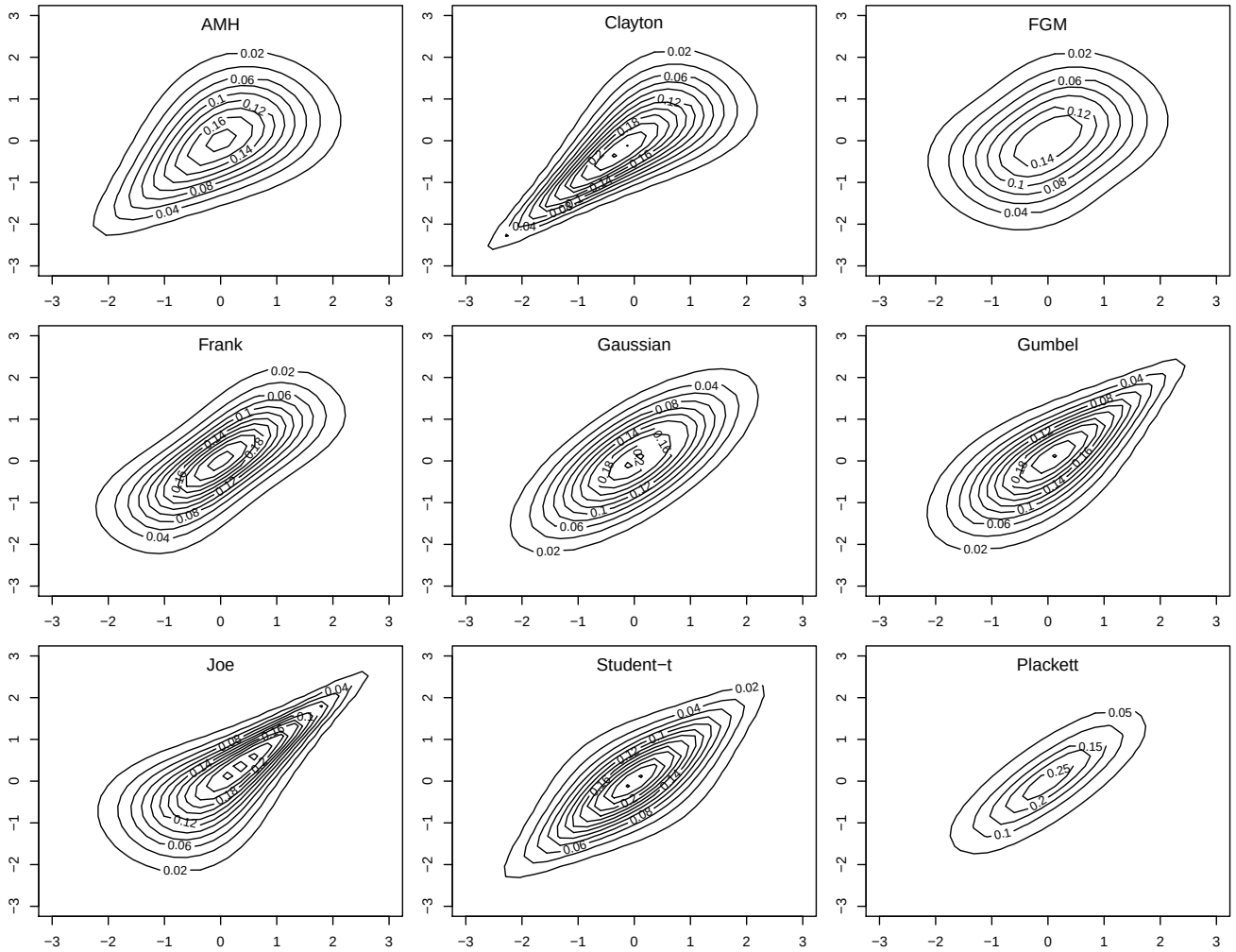
$P(T_{1i} > t_{1i}, T_{2i} > t_{2i}|\mathbf{x}_i; \boldsymbol{\delta})$ can be expressed as

$$S(t_{1i}, t_{2i}|\mathbf{x}_i; \boldsymbol{\delta}) = C\left(S_1(t_{1i}|\mathbf{x}_{1i}; \boldsymbol{\beta}_1), S_2(t_{2i}|\mathbf{x}_{2i}; \boldsymbol{\beta}_2); m\{\eta_{3i}(\mathbf{x}_{3i}; \boldsymbol{\beta}_3)\}\right),$$

where S_1 and S_2 are conditional marginal survival functions, C is a copula function with coefficient $\theta_i = m\{\eta_{3i}(\mathbf{x}_{3i}; \boldsymbol{\beta}_3)\}$ capturing the conditional dependence of (T_{1i}, T_{2i}) across observations, m is an inverse link function, and $\eta_{3i}(\mathbf{x}_{3i}; \boldsymbol{\beta}_3)$ is a predictor which includes generic covariate effects.

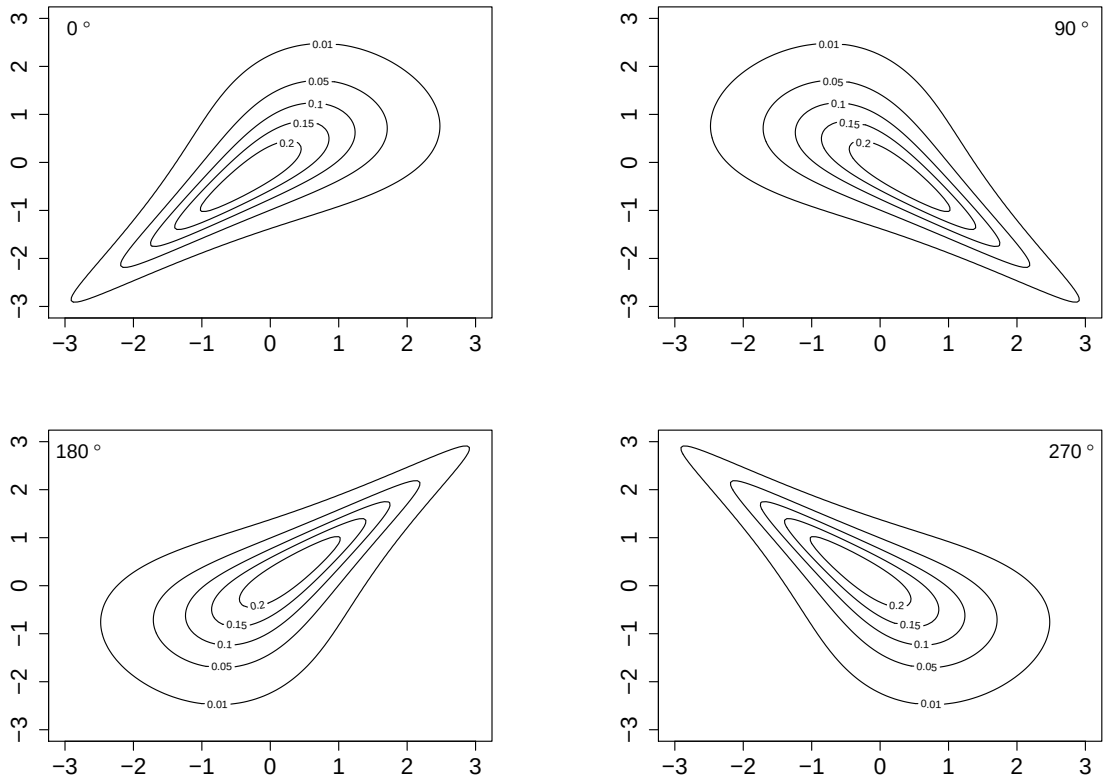
The margins are modelled as $S_v(t_{vi}|\mathbf{x}_{vi}; \boldsymbol{\beta}_v) = G_v\{\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)\}$ $(v = 1, 2)$, where G_v is an inverse link function.

Some of the copulae implemented in GJRM



So we do not have to necessarily assume Gaussian dependence.

Rotated versions of Clayton and mixed copulae



Positive and negative tail dependencies can be simultaneously modelled using a switching model. For instance, $\zeta\mathcal{C}(u,v) + (1-\zeta)\mathcal{C}_{90}(u,v)$ where ζ is a binary switching variable.

Additive predictor

As opposed to $\eta_{3i}(\mathbf{x}_{3i};\boldsymbol{\beta}_3)$, $\eta_{vi}(t_{vi},\mathbf{x}_{vi};\boldsymbol{\beta}_v)$ $(v=1,2)$ must include baseline functions of time. But t_{vi} can be treated just like a covariate. Therefore, we consider a generic predictor η_i and an overall covariate vector called $\mathbf{z}_i$ which is made up of t_i and $\mathbf{x}_i$ and define

$$\eta_i = \beta_0 + s_1(\mathbf{z}_{1i}) + s_2(\mathbf{z}_{2i}) + \ldots + s_K(\mathbf{z}_{Ki}), \quad i = 1,\ldots,n.$$

For each i the generic $s(\mathbf{z}_i)$ can be approximated as $\sum_{j=1}^{J}\beta_j b_j(\mathbf{z}_i)$.

For all observations, we have $\mathbf{Z}\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_J)^{\mathsf{T}}$ and $Z[i,j] = b_j(\mathbf{z}_i)$. Hence,

$$\boldsymbol{\eta} = \beta_0\mathbf{1}_n + \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_K\boldsymbol{\beta}_K.$$

Quadratic penalty $\lambda_k\boldsymbol{\beta}_k^{\mathsf{T}}\mathbf{S}_k\boldsymbol{\beta}_k$ imposes specific properties on the k^{th} function.

Key ingredients are $\mathbf{Z}_k$ and $\mathbf{S}_k$.

Linear and non-linear effects

For binary or categorical predictors, $s_k(\mathbf{z}_{ki})$ is approximated by

$$\mathbf{z}_{ki}^{\mathsf{T}}\boldsymbol{\beta}_k.$$

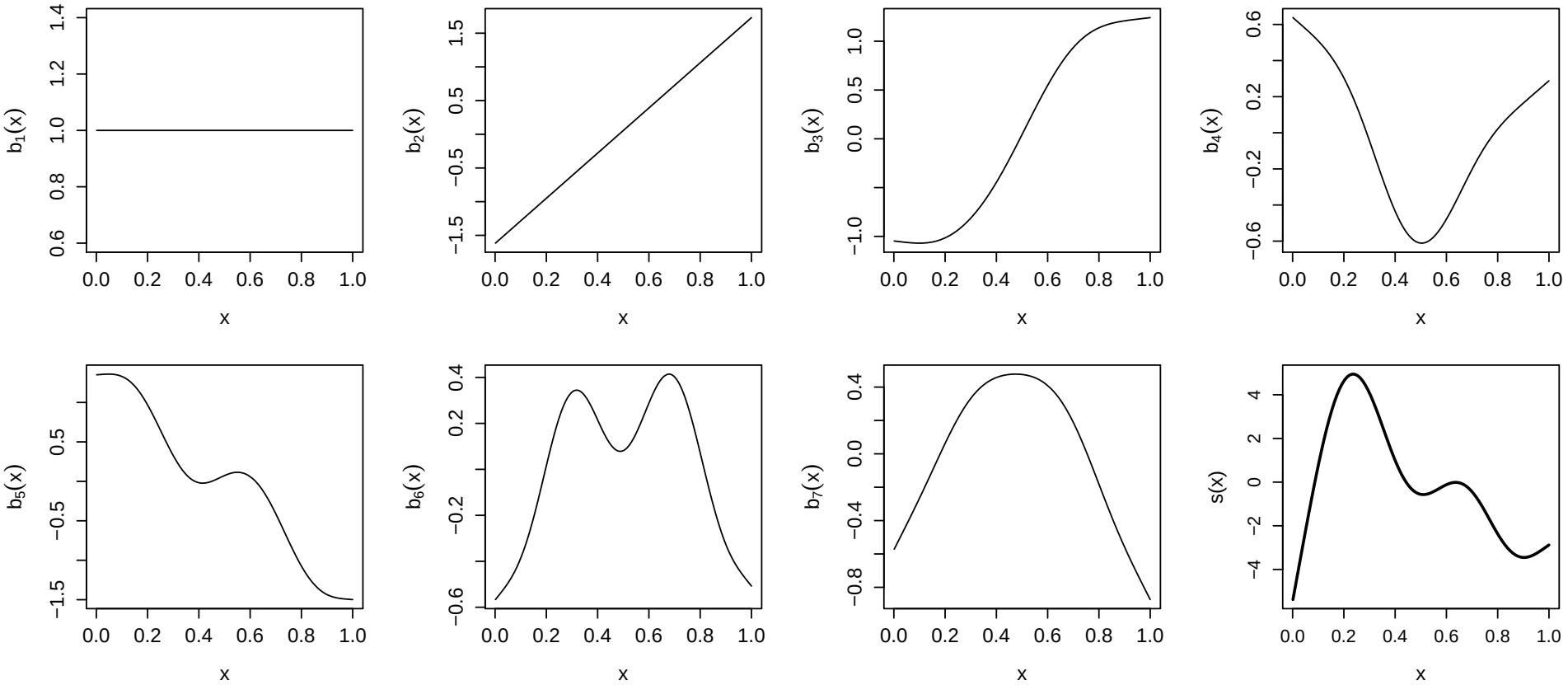
$\mathbf{Z}_k$ is obtained by stacking all covariate vectors. Typically, $\mathbf{S}_k = \mathbf{0}$ but for the interviewer parameters $\mathbf{S}_k = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. (This has the interpretation of a *random effect*.)

For continuous variables,

$$\sum_{j_k=1}^{J_k} \beta_{kj_k} b_{kj_k}(z_{ki}),$$

where the $b_{kj_k}(z_{ki})$ are known spline basis functions. $\mathbf{Z}_k$ comprises the basis function evaluations for each i. $\mathbf{S}_k = \int \mathbf{d}_k(z_k)\mathbf{d}_k(z_k)^{\mathsf{T}} dz_k$, where the j_k^{th} element of $\mathbf{d}_k(z_k)$ is given by $\partial^2 b_{kj_k}(z_k)/\partial z_k^2$.

Thin plate regression spline example



Spatial effects

Geographic location of respondents is exploited using

$$\mathbf{z}_{ki}^{\mathsf{T}}\boldsymbol{\beta}_k,$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kR})^{\mathsf{T}}$ contains R spatial effects and

$$\mathbf{Z}_k[i, r] = \begin{cases} 1 & \text{if the observation belongs to region } r \\ 0 & \text{otherwise} \end{cases}, \quad r = 1, \ldots, R.$$

The smoothing penalty is an adjacency matrix

$$\mathbf{S}_k[r, q] = \begin{cases} -1 & \text{if } r \neq q \wedge r \text{ and } q \text{ are adjacent neighbors} \\ 0 & \text{if } r \neq q \wedge r \text{ and } q \text{ are not adjacent neighbors} \\ N_r & \text{if } r = q \end{cases},$$

where N_r is the total number of neighbors for region r. (This has the interpretation of a *Gaussian Markov random field*.)

Function G

Inverse link function $G(\eta)$ can be specified as $\exp\{-\exp(\eta)\}$ and $\frac{\exp(-\eta)}{1+\exp(-\eta)}$.

The respective link functions, $g(S)$, are then $\log\{-\log(S)\}$ and $-\log\left(\frac{S}{1-S}\right)$.

Consider the marginal model

$$g\{S(t_i|\mathbf{x}_i)\} = g\{S_0(t_i)\} + \sum_{k=1}^{K} s_k(\mathbf{x}_{ki}),$$

where $S_0(t_i)$ is a background survival function. This model yields the proportional hazards model when choosing the log-log link. In fact,

$$\log\{H(t_i|\mathbf{x}_i)\} = \log\{H_0(t_i)\} + \sum_{k=1}^{K} s_k(\mathbf{x}_{ki}),$$

where $H(t_i|\mathbf{x}_i) = -\log\{S(t_i|\mathbf{x}_i)\}$, and $H_0(t_i) = -\log\{S_0(t_i)\}$ is the cumulative background hazard function. $g\{S_0(t_i)\}$ is represented using $s_0(t_i)$.

Penalised likelihood-based inference

The likelihood function involves terms like $\partial \eta_{vi}/\partial y_{vi}$ $(v = 1, 2)$ which *must* be positive. Assume that no covariates are available and let $s(y_i) = \sum_{j=1}^{J} \gamma_j b_j(y_i)$. We have $s'(y_i) \geq 0$ if $\gamma_j \geq \gamma_{j-1}, \forall j$. This is achieved by $\gamma_1 = \beta_1, \gamma_j = \beta_1 + \sum_{j=2}^{J} \exp(\beta_j), \forall j = 2, ..., J$. The penalty term is set up to penalise the squared differences between adjacent β_j, starting from β_2.

Because of right-censoring, the score vector and Hessian matrix have complicated structures, which are considerably further complicated by the non-linear dependence of $\boldsymbol{\gamma}$ on $\boldsymbol{\beta}$.

Main computational challenges are: (i) provide a stable and fast algorithm for problems which may not be concave and/or exhibit regions that are close to flat, (ii) estimate $\boldsymbol{\lambda}$ using an automatic approach.

Cont'd

Estimation is based on direct optimisation of

$$\ell_p(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}) - \frac{1}{2}\boldsymbol{\psi}^{\mathsf{T}}\left(\sum_k \lambda_k \mathbf{S}_k\right)\boldsymbol{\psi}.$$

For a fixed value of $\boldsymbol{\lambda}$, estimation of $\boldsymbol{\psi}$ is achieved via

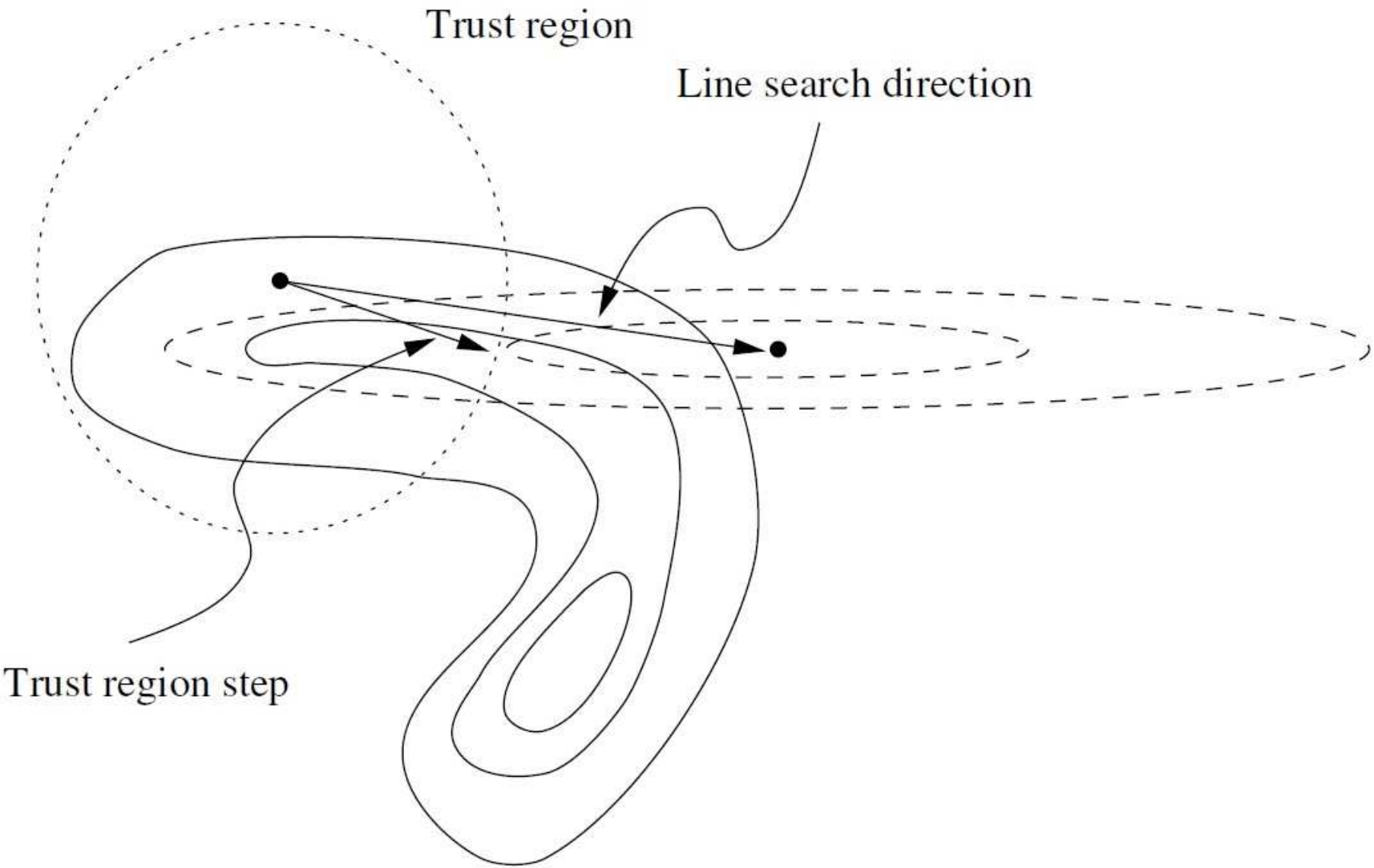
$$\min_{\mathbf{p}} \ \breve{\ell}_p(\boldsymbol{\psi}^{[a]}) \stackrel{\text{def}}{=} -\left\{\ell_p(\boldsymbol{\psi}^{[a]}) + \mathbf{p}^{\mathsf{T}}\mathbf{g}_p^{[a]} + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\boldsymbol{\mathcal{H}}_p^{[a]}\mathbf{p}\right\} \quad \text{so that} \quad \|\mathbf{p}\| \leq r^{[a]},$$

$$\boldsymbol{\psi}^{[a+1]} = \boldsymbol{\psi}^{[a]} + \arg\min_{\mathbf{p}} \ \breve{\ell}_p(\boldsymbol{\psi}^{[a]}),$$

where $\mathbf{g}_p$ and $\boldsymbol{\mathcal{H}}_p$ are the penalised score and Hessian. $\breve{\ell}_p(\boldsymbol{\psi}^{[a]})$ is minimized subject to the constraint that the solution falls within a trust region with radius $r^{[a]}$.

Accept or reject and region expanded or shrunken are based on the ratio between the improvement in the objective function when going from $\boldsymbol{\psi}^{[a]}$ to $\boldsymbol{\psi}^{[a+1]}$ and that predicted by the approximation.

A pictorial representation



Automatic multiple smoothing parameter selection

After some manipulation, the estimator for $\boldsymbol{\psi}$ can be expressed as

$$\boldsymbol{\psi}^{[a+1]} = \left(-\boldsymbol{\mathcal{H}}^{[a]} + \mathbf{S}\right)^{-1} \sqrt{-\boldsymbol{\mathcal{H}}^{[a]}}\mathbf{M}^{[a]},$$

where $\mathbf{M}^{[a]} = \sqrt{-\boldsymbol{\mathcal{H}}^{[a]}}\boldsymbol{\psi}^{[a]} + \sqrt{-\boldsymbol{\mathcal{H}}^{[a]}}^{-1}\mathbf{g}^{[a]}$. Prediction for $\mathbf{M}$ is $\hat{\mu}_{\mathbf{M}} = \sqrt{-\boldsymbol{\mathcal{H}}}\hat{\boldsymbol{\psi}} = \mathbf{AM}$, where $\mathbf{A} = \sqrt{-\boldsymbol{\mathcal{H}}}\left(-\boldsymbol{\mathcal{H}} + \mathbf{S}\right)^{-1}\sqrt{-\boldsymbol{\mathcal{H}}}$.

Aim: estimate $\boldsymbol{\lambda}$ so that $\hat{\mu}_{\mathbf{M}}$ is as close as possible to $\mu_{\mathbf{M}}$ ...

$$\boldsymbol{\lambda}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}} \; \|\mathbf{M}^{[a+1]} - \mathbf{A}^{[a+1]}\mathbf{M}^{[a+1]}\|^2 + 2\mathrm{tr}(\mathbf{A}^{[a+1]}).$$

Key ingredients are obtained as a byproduct of the estimation step for $\boldsymbol{\psi}$. Also, computations are more stable because we are using $\mathbf{g}$ and $\boldsymbol{\mathcal{H}}$ as a whole. (It can be proved that this is approximately equivalent to the AIC.)

Bayesian 'confidence' intervals

Well-calibrated intervals for linear and non-linear functions of the model's coefficients are obtained using $\psi \dot{\sim} \mathcal{N}(\hat{\psi}, -\widehat{\mathcal{H}}_p^{-1})$. (This result comes from using the distribution of $\mathbf{M}$, making the large sample assumption that $\mathcal{H}$ can be treated as fixed, and making the usual Bayesian assumption on the prior of ψ for smooth models.)

An important advantage of this result is that intervals for non-linear functions of ψ can be conveniently obtained by simulation, hence avoiding computationally expensive parametric bootstrap.

Another advantage is that the distribution of non-linear functions of ψ need not be symmetric.

Some other details

Large sample properties can be derived. For instance, assuming that the number of spline bases is large 'enough' and under some customary assumptions it can be proved that $\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^0 = O_P(1/\sqrt{n})$

Model building can be aided using the AIC, BIC, Cox-Snell residuals and hypothesis testing.

library(GJRM)

An example of call is

```
eq1 <- y1 ~ x1 + s(x2, bs = "tp") + s(y1, bs = "mpi")
eq2 <- y2 ~ x1 + s(y2, bs = "mpi")
eq3 <-    ~ s(x2)
f.list <- list(eq1, eq2, eq3)

out <- copulaReg(f.list, data = mydata, surv = TRUE,
                 BivD = "PL", margins = c("PH", "PO"),
                 cens1 = u1, cens2 = u2)
```

where `bs` specifies the spline basis type (e.g., `tp` for thin plate regression spline (the default) and `mpi` for monotonic P-spline). Monotonic P-splines must always be used for smooth terms of the responses. Post-estimation functions include `summary()`, `plot()`, `conv.check()`, `post.check()` and `hazsurv.plot()`.

Estimating the dependence between colon cancer recurrence and death

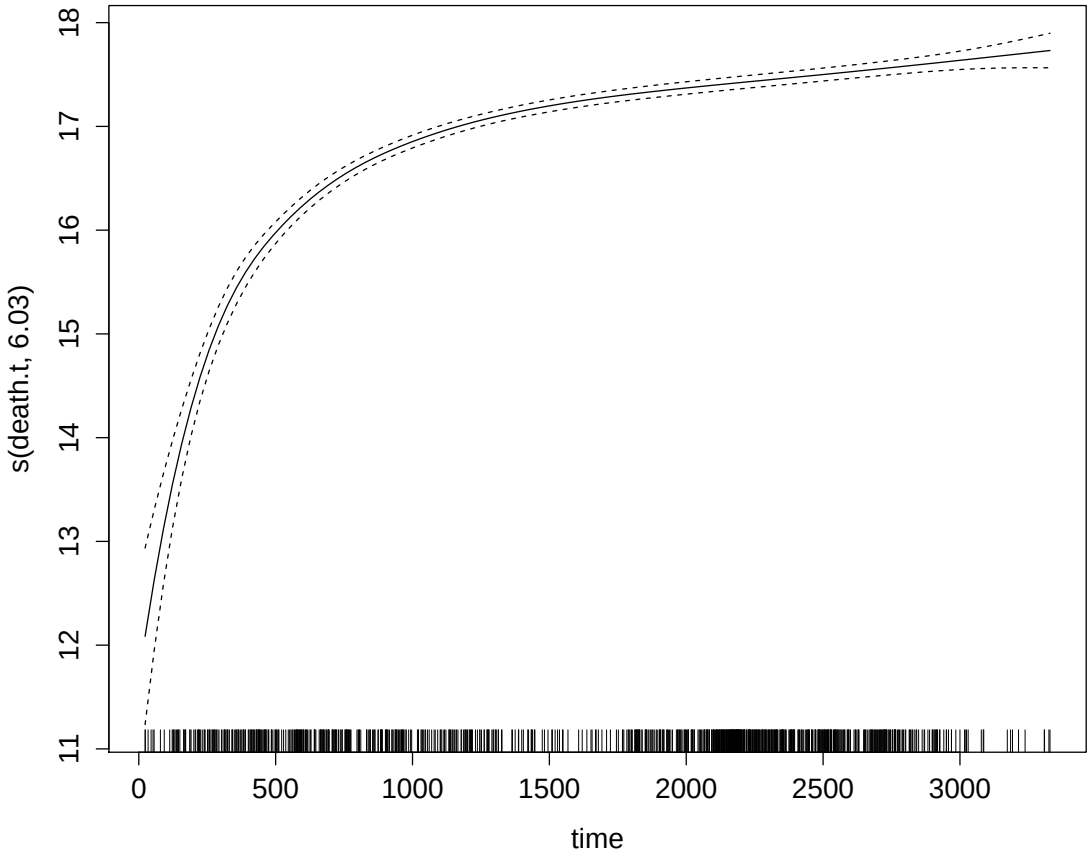
The data are from a trial of adjuvant chemotherapy for colon cancer, where levamisole and fluorouracil are used to treat cancer. The treatments could not be blinded, so no placebos were used. There are around 15 variables. A model is

```
eq1   <- rectime ~ s(rectime, bs = "mpi") + treat + s(age) + adhere + obstruct + ..
eq2   <- death.t ~ s(death.t, bs = "mpi") + treat + s(age) + adhere + obstruct + ..
eq3   <-         ~                          treat + s(age) + perfor + differ
```

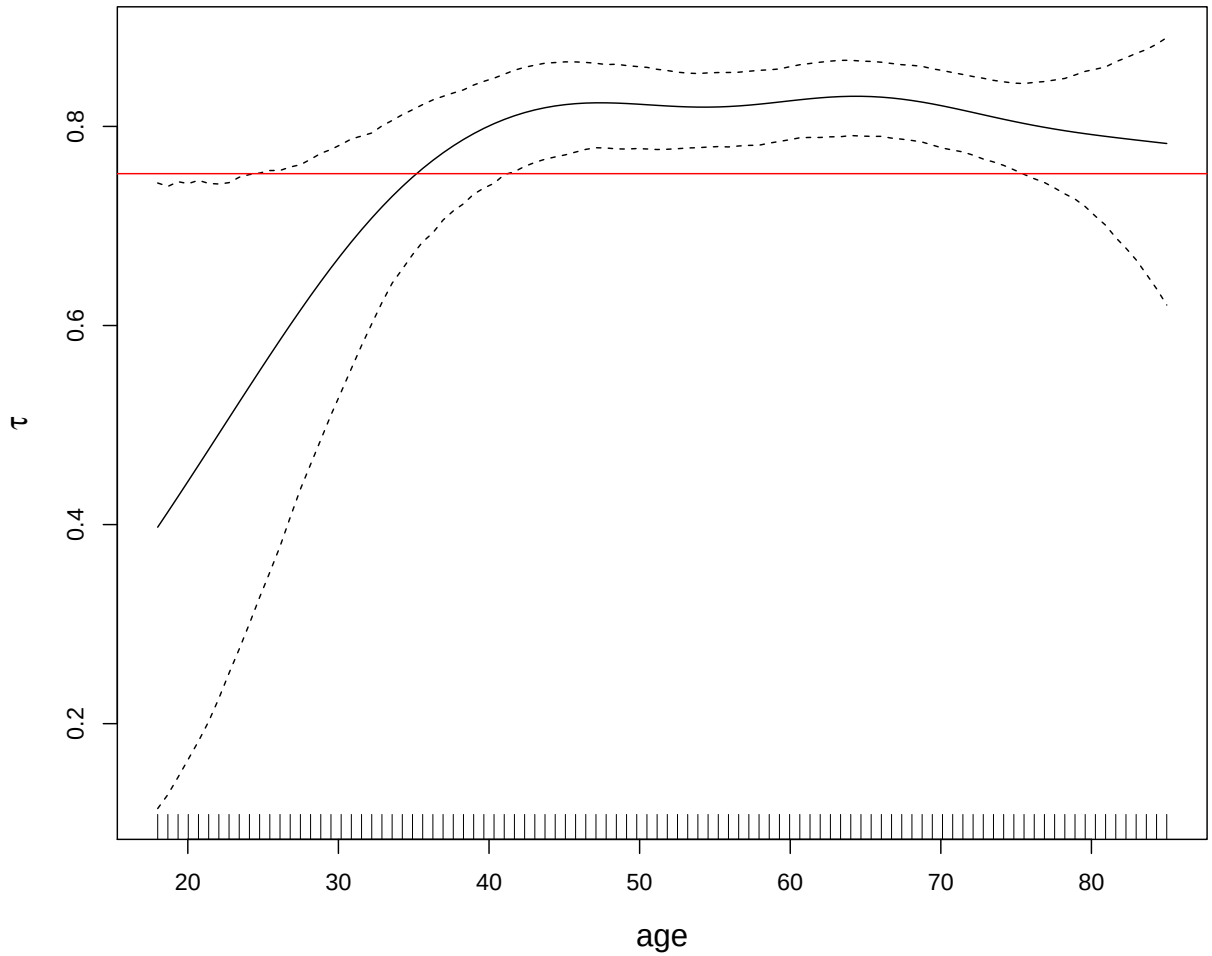
where `rectime` is the time to recurrence with censoring indicator `event1`, `death.t` is the time to death with censoring indicator `event2`, `treat` is made up of three categories (observation, levamisole, levamisole + fluorouracil), `obstruct` (obstruction of colon by tumour), `perfor` (perforation of colon), `adhere` (adherence to nearby organs), etc ... and `BivD = "J180", margins = c("PH", "PH")`.

Some results

The individuals taking levamisole and fluorouracil have a colon cancer recurrence (mortality) rate that is about 38% (30%) smaller than that of people taking nothing or levamisole alone.



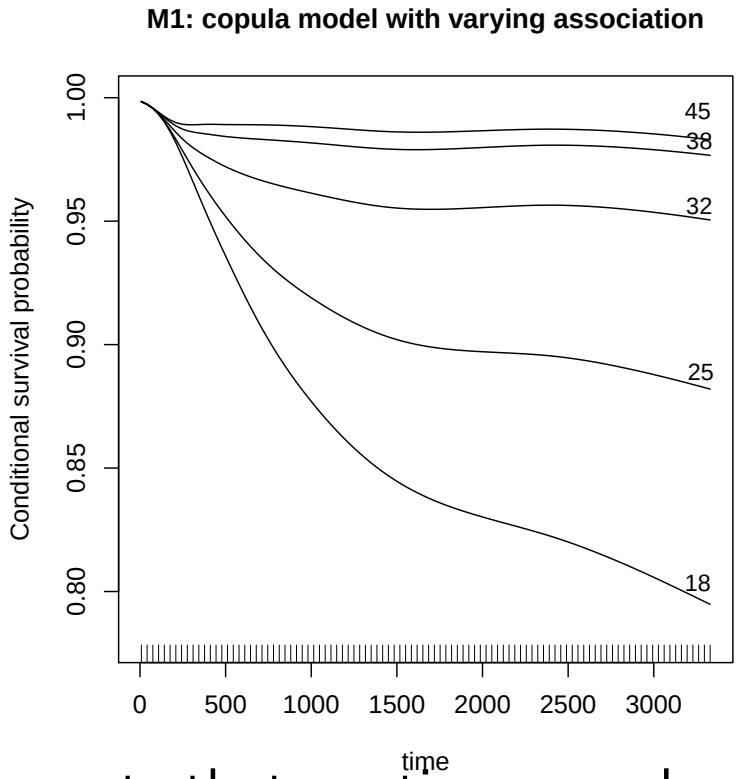
Impact of age on Kendall's τ

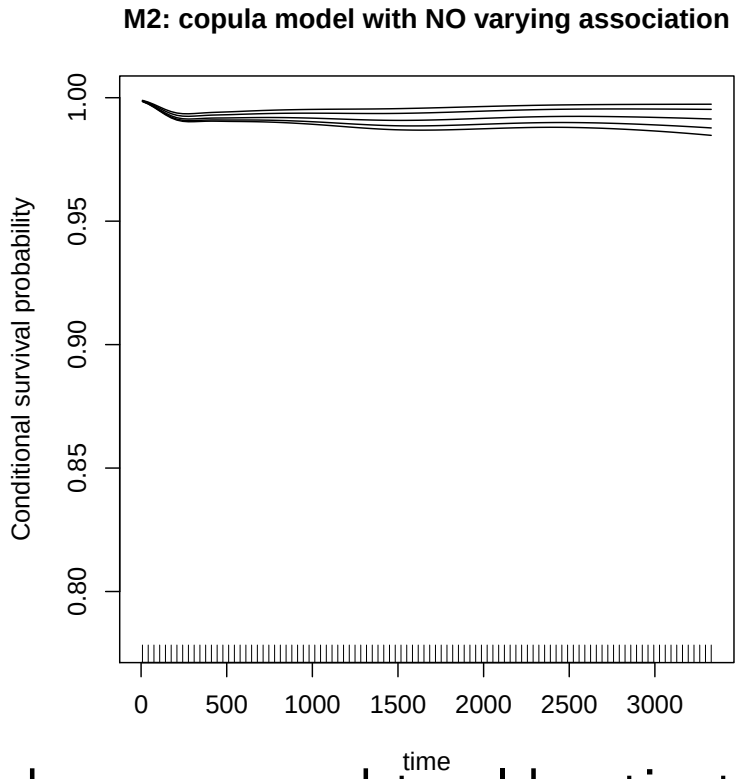


This suggests non-constant association between time to recurrence and death and that stronger concomitance of the two events is present among patients who are at least 40.

Some survival plots

The plots below show the probability of surviving given that the cancer did not recur by age group.





M1 suggests that, as time goes by and as compared to old patients, young patients are less likely to survive if the cancer did not recur. This finding is not conveyed by M2. Early detection of asymptomatic recurrences is especially relevant for older patients and could be taken into account in surveillance guidelines.

Some future work

Copula models with mixed binary, discrete, continuous and survival margins.

Extensions to more than two dimensional settings.

Assess empirical performance of functional predictors in the survival context and beyond.

Extend the set of link functions available.

Apply the models to different case studies.