INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0402

LOCAL DEPENDENCE ESTIMATION USING NON-PARAMETRIC ARCHIMEDEAN COPULAS

F. VANDENHENDE, P. LAMBERT

http://www.stat.ucl.ac.be

Local dependence estimation using non-parametric Archimedean copulas

François Vandenhende[1,2] and Philippe Lambert[2,3]

[1] Statistics and Information Sciences, Eli Lilly & Company.
[2] Institute of Statistics, Université Catholique de Louvain.
[3]Unité d'épidémiologie, biostatistique et méthodes opérationnelles,
Faculté de Médecine, Université catholique de Louvain, Belgium.

SUMMARY. The selection of a dependence model is often arbitrary but rarely automated when modeling multivariate data. We define new non-parametric copulas that overcome the problem. We focus our attention to *Archimedean* copulas. Then, the non-parametric family is obtained using a continuous piecewise log-linear combination of existing Archimedean generators. We provide an efficient constrained least-squares method to estimate the involved coefficients. We show how these coefficients can be interpreted as local dependence measures. We also present a penalization algorithm for smoothing that dependence over the joint distribution. The usefulness of the non-parametric family is illustrated in a example: the lifetime analysis of Danish twins, where dependence increases with decreasing joint survival probability.
keywords: Archimedean copula, local dependence, quantile function, non-parametric smoothing.

1. Introduction

The Archimedean copula family (Nelsen, 1999, pp. 89–124) is an important class of dependence functions which can be used to generate joint distributions of random variables X_1 and X_2 with specified marginal distributions $F_{X_1}(x_1) = \Pr(X_1 \leq x_1)$ and $F_{X_2}(x_2) = \Pr(X_2 \leq x_2)$. Members C of this class are bivariate distributions with uniform margins on $[0,1]^2$ constructed from a continuous, strictly decreasing, and convex function ϕ such that

$$C(u,v) = \phi^{[-1]}\{\phi(u) + \phi(v)\}. \tag{1}$$

ϕ is called the *generator* of the copula. It is defined on $(0,1]$ such that $\phi(1) = 0$. When $\phi(0) = +\infty$, we say that the generator is strict. In that case, $\phi^{[-1]}$ is the inverse of ϕ. With non-strict generators (when $\phi(0)$ is finite), $\phi^{[-1]}$ is a pseudo-inverse of ϕ, equal to ϕ^{-1} on $(0, \phi(0)]$ and to zero, otherwise.

---

email: francois@lilly.com. Correspondence to: Francois Vandenhende, Eli Lilly and Company, 11 rue Granbompre, B-1348 Mont-Saint-Guibert, Belgium. Phone: +32-10476469; Fax: +32-10476475.

Then,

$$H(x_1, x_2) = C\{F_{X_1}(x_1), F_{X_2}(x_2)\}$$

is a joint distribution of (X_1, X_2). The name *copula* comes from the fact that C combines the marginal distributions to create H. The strength of dependence between X_1 and X_2 is a function of the shape of C.

Genest and Rivest (1993) derived the distribution function

$$K(p) = \Pr\{C(U, V) \leq p)$$

of an Archimedean copula C with uniform rv's U and V. For any $p \in [0, 1]$, it is equal to

$$K(p) = p - \frac{\phi(p)}{\phi'(p^+)},$$

where $\phi'(p^+)$ denotes the derivative to the right of ϕ at p. One rank-based dependence measure, Kendall's tau (Kruskal, 1958), is derived from that distribution. When X_1 and X_2 are absolutely continuous, Genest and Mackay (1986) showed that

$$\begin{aligned} \tau &= 4E_K(p) - 1 \\ &= 1 + 4\int_0^1 \frac{\phi(p)}{\phi'(p^+)} dp, \end{aligned} \tag{2}$$

where $E_K(p)$ is the expected value of $C(U, V)$. That global concordance measure has a simple interpretation. It estimates the difference between the probabilities of concordance and discordance in (X_1, X_2). With Archimedean copulas, τ is a direct function of the generator ϕ.

The class of Archimedean copulas encompasses many well-known bivariate parametric distributions, such as the families of Frank, Clayton or Gumbel (see Table 1). A more complete list is given in Nelsen (1999, pp. 94–97). For these three families, the generator ϕ is a derivable function having a single parameter θ that estimates the sign and the strength of the dependence. In every case, one particular or a limiting value for θ leads to the generator $\phi(p) = -\log(p)$ that creates independence, as then, $C = \exp\{\log(U) + \log(V)\} = UV$ is the product of the margins U and V. For each of these families, there is a 1 to 1 mapping between θ and τ, as illustrated in Table 1.

In practice, once a family has been selected, one can estimate the dependence parameter by equating the sample-based estimate of Kendall's tau to its population version of Table 1. As τ is function of the first moment of K (see Eq. 2), the estimation can be seen as a moment-based method. Other methods such as

Table 1

Selected bivariate Archimedean copulas having a single dependence parameter (θ).

Name $\theta \in$	$\phi_\theta(p)$	$C_\theta(u,v)$	Kendall's tau
Frank $(-\infty,\infty)/\{0\}$	$-\log \frac{e^{-\theta p}-1}{e^{-\theta}-1}$	$-\frac{1}{\theta}\log[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}]$	$1-\frac{4}{\theta}\{1-D^*(\theta)\}$
Clayton $[-1,\infty)/\{0\}$	$\frac{1}{\theta}(p^{-\theta}-1)$	$\max\{(u^{-\theta}+v^{-\theta}-1)^{-1/\theta},0\}$	$\frac{\theta}{\theta+2}$
Gumbel $[1,\infty)$	$(-\log p)^\theta$	$\exp[-\{(-\log u)^\theta+(-\log v)^\theta\}^{1/\theta}]$	$\frac{\theta-1}{\theta}$

*: $D(x)=\frac{1}{x}\int_0^x \frac{t}{e^t-1}dt$

maximum likelihood (ML) are also possible. See, for instance, Vandenhende and Lambert (2000) for an application.

The selection of an Archimedean family is often arbitrary and rarely automated. Genest and Rivest (1993) provided a graphical method allowing the comparison of different parametric distributions K with the empirical counterpart. In Vandenhende and Lambert (2000), copula selection was done using Akaike's information criteria from the ML fit of several copula models having the same marginal components. In the two cases, only a limited set of families are investigated.

In this paper, we overcome this selection problem by defining flexible non-parametric Archimedean copulas. The distribution K is estimated using local polynomials. The involved parameters quantify dependence. The layout is as follows. In Section 2, we transform the generator into a linear function of dependence parameters. The result is a univariate quantile function. The relationship between quantile functions and Archimedean copula is also studied in that section. In Section 3, we create a non-parametric version of the generator from the back-transformation of local polynomials. Dependence properties of parameters are discussed in Section 4. A least-squares estimation method is presented in Section 5. The practical usefulness of the non-parametric copula is illustrated in Section 6. Section 7 contains a discussion of key results.

2. Archimedean copula and quantile functions

2.1 *Deriving quantile functions from Archimedean copulas*

We start our developments by a simple log-transformation

$$Q(p) = -\log\{\phi(p)\} \tag{3}$$

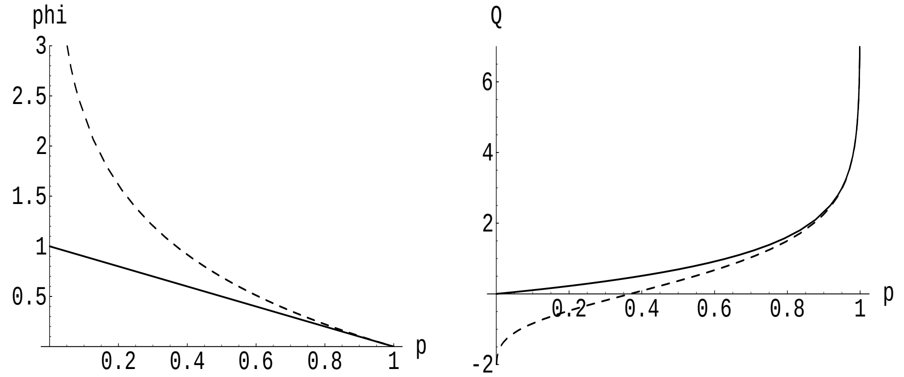

Figure 1. The transformation $Q = -\log(\phi)$ of a Archimedean generator creates a continuous univariate quantile function. The range of Q is unbounded in $\mathbb{R}$ with strict generators (solid line). It is bounded to the left with non-strict generators (dashed line).

of the generator ϕ. As shown in Figure 1, Q is a continuous and increasing function on $(0, 1)$. It tends to $+\infty$ when $p \to 1$. With strict generators, $\lim_{p\to 0} Q(p) = -\infty$, otherwise $Q(0)$ is finite.

The function Q that maps $(0, 1)$ into $\mathbb{R}$ is a *quantile function*: it is the inverse of a univariate cumulative distribution, for a continuous random variable $Z \in \mathbb{R}$. This means that, using the transformation (3), every Archimedean copula family is associated to a univariate distribution $F = Q^{-1}$. With strict generators, the domain of Z is unbounded in $\mathbb{R}$. With non-strict generators, it is bounded to the left, producing, for example, distribution functions for positive random variables.

One can thus use (3) to derive univariate distributions from known Archimedean copula families. For example, Frank's family (see Table 1) induces the density

$$f(z) = \frac{e^{-z}(e^{\theta} - 1)}{\theta + e^{\theta}(\exp(e^{-z}) - 1)\theta},$$

for $z \in \mathbb{R}$. This example illustrates one of the many original univariate distributions that can be derived from Archimedean generators. Given the wide variety of Archimedean copulas, a large number of univariate distributions could be (re)-discovered using the simple transformation (3) of their generator. The study of properties of so-derived distributions is ongoing.

One sub-family of generators is such that

$$\phi(p; \theta) = \{\phi(p; \theta_0)\}^{\theta}.$$

Then, the corresponding quantile function is

$$Q(p;\theta) = \theta S(p), \tag{4}$$

with $S(p) = \exp\{-\phi(p;\theta_0)\}$, coming from some reference generator . The dependence parameter θ appears as a linear factor in (4). Three main copula families have that representation. They lead to very well known distribution functions:

1. the generator of Gumbel's family (see Table 1) leads to the extreme-value (EV) distribution $F(z) = \exp\{-\exp(-\frac{z}{\theta})\}$ for $z \in \mathbb{R}$,
2. the family generated by $\phi = (\frac{1}{p} - 1)^\theta$, $\theta \geq 1$ is associated to the logistic distribution $F(z) = \frac{1}{1+\exp(-z/\theta)}$ for $z \in \mathbb{R}$, and
3. the generator $\phi(t) = (1-t)^\theta$, with $\theta \geq 1$ is related to an exponential distribution $F(z) = 1 - \exp(z/\theta)$ for $z \in \mathbb{R}^+$.

In (4), the dependence parameter θ appears as a scale parameter in the associated quantile function. Indeed, when the variance v_S under S exists, the variance under Q is $v_Q = \theta v_S$.

In addition to the scale, a location parameter $c \in \mathbb{R}$ can also be added to the quantile function, producing

$$Q(p) = c + \theta S(p),$$

a flexible univariate distribution defined from any Archimedean generator $\phi = \exp(-S)$. The following theorem summarizes these results.

Theorem 1: Let $Q(p) = -\log\{\phi(p)\}$, $p \in (0,1)$ where ϕ is an Archimedean generator function. For any $c \in \mathbb{R}$ and $\theta > 0$, let $Q'(p) = c + \theta Q(p)$. Then,

- Q and Q' are valid univariate quantile functions for a continuous random variable Z in $\mathbb{R}$.
- If ϕ is strict, the domain of Z is unbounded in $\mathbb{R}$, otherwise it is bounded to the left.
- If the mean μ_Q and variance v_Q from Q exist, the mean and variance derived from Q' are equal to $\mu_{Q'} = c + \theta\mu_Q$ and $v_{Q'} = \theta^2 v_Q$, respectively.

2.2 *Constructing Archimedean copulas from quantile functions*

In the previous section, we showed that to any Archimedean copula corresponds a univariate quantile function. The opposite is not necessarily true. Indeed, selected properties need to be verified by univariate quantile functions Q to ensure that $\exp(-Q)$ generates an Archimedean copula. We define those conditions in this section.

Let Q be a quantile function for a continuous random variable $Z \in \mathbb{R}$: $Q(p) : [0,1] \to \mathbb{R}$. As Q is continuous and increasing, $\exp(-Q)$ will generate a valid Archimedean copula if

- Q(1)=$+\infty$: The domain of Z is unbounded to the right.
- $\exp(-Q)$ is convex.

To express the convexity condition, we define, when Q is derivable, the quantile density associated to Q by $q(p) = \frac{dQ(p)}{dp}$. The inverse of the quantile density, which corresponds to the density f of Z at the percentile p is denoted by $f_Q(p) = f(Q(p)) = \frac{1}{q(p)}$. The quantile density $f_Q(p)$ is non-negative on $[0,1]$, with $f_Q(1) = 0$ when Z is not bounded to the right. When Z is also unbounded to the left, taking every possible value in $\mathbb{R}$, $f_Q(0) = 0$. Otherwise $f_Q(0) > 0$.

When Q is twice differentiable, the convexity condition becomes

$$\exp\{-Q(p)\}'' \geq 0.$$

It can be reformulated as

$$\begin{aligned} f'_Q(p) &\geq -1, \forall p \in [0,1] \\ \text{or} \quad \frac{\partial \log(f(z))}{\partial z} &\geq -1, \forall z. \end{aligned} \tag{5}$$

One can thus verify if (5) holds for any standard univariate continuous distribution unbounded to the right to see if it generates an Archimedean copula. For example, let us consider the extreme-value distribution with scale parameter θ. As its entire domain is $\mathbb{R}$, it is not bounded to the right. Since, the log-density has first derivative always superior to $-1/\theta$, condition (5) is verified when $\theta \geq 1$. In that case, it generates Gumbel's family from Table 1 with dependence parameter θ.

One counter example is the normal distribution having mean c and standard deviation θ. The derivative of its log-density is $-(z-c)/\theta$, which tends to $-\infty$ as $z \to +\infty$, whatever c and θ. Thus one cannot create an Archimedean copula using the univariate normal quantile function.

2.3 *Families of generators*

In this section we enrich the class of Archimedean copulas using linear operations on quantile functions that meet the criteria for generating Archimedean copulas. Let us define Ω, the class of such quantile functions $S : [0,1] \to \mathbb{R}$ twice differentiable that verify $S(1) = +\infty$ and $f'_S(p) \geq -1$. Following Gilchrist (2000), we present a series of linear operations γ on S so that $Q = \gamma(S)$ is also in Ω.

The list is as follows:

- Translation: If $S \in \Omega$ and $c \in \mathbb{R}$, then $Q = c + S \in \Omega$. In addition, Q and S define the same copula

$$C(u,v) = S^{[-1]}(-\log[\exp\{-S(u)\} + \exp\{-S(v)\}]), \tag{6}$$

  in which $S^{[-1]} = S^{-1}$ on the range of S and is zero otherwise. For non-strict generators, we can thus set the lower bound $S(0) = 0$, arbitrarily.

- Addition: If $S_1, S_2 \in \Omega$, then $Q = S_1 + S_2 \in \Omega$.
- Linear Interpolation: If $S_1, S_2 \in \Omega$ and $\theta \in [0,1]$, then $(1-\theta)S_1 + \theta S_2 \in \Omega$
- Scaling: If $S \in \Omega$ and $\theta \geq 1$, then $\theta S \in \Omega$.

According to the Translation properties, the resulting copula is unchanged when changing the location (mean) of the related quantile function. The Scaling operation generates the so-called β-family of generators. See Nelsen (1999, p. 114) for properties of that family. The condition $\theta \geq 1$ is required to insure the convexity of the resulting generator.

Any combination of these operations also remains in Ω. Thus, one can construct new generators Q for Archimedean copulas as linear combinations of generators S_i in Ω. The resulting copula will have a closed-form expression provided that Q^{-1} has a closed form expression. When distinct families S_i are combined into Q, the inverse Q^{-1} has usually no closed analytical form. Numerical procedures are then required to evaluate the generated copula C. We shall see later that a direct evaluation is not required to obtain least-squares estimates for the parameters θ_i. Only the equation for Q is required. However this is not true when MLEs are requested.

3. Non-parametric Archimedean copulas

Instead of combining several quantile functions $S_i \in \Omega$ $i = 0, ..., n$ on the whole probability range $[0,1]$, one can use a distinct function S_i for every interval

$[t_i, t_{i+1})$ of a partition of $[0, 1]$ and create a continuous quantile function Q using local linear combinations of the S_i. Under some conditions, the result is such that $\exp(-Q)$ generates an Archimedean copula. That copula is indeed a non-parametric (or highly parameterized) Archimedean copula, having $n+1$ parameters that quantify the dependence in the corresponding intervals of the partition. In this section we detail how to accomplish this.

We select arbitrarily n knots in the $(0, 1)$ interval, such that $0 < t_1 < ... < t_n < 1$. Two additional knots are defined at $t_0 = 0$ and $t_{n+1} > 1$, so that $1 \in [t_n, t_{n+1})$. The suggested piecewise interpolating function Q is defined using (n+1) quantile functions $S_i \in \Omega$ as

$$Q(p) = \sum_{i=0}^{n} (\theta_{0,i} + \theta_i S_i(p)) I(t_i \leq p < t_{i+1}). \tag{7}$$

The following theorem summarizes conditions under which $\exp(-Q)$ generates an Archimedean copula and some properties of the resulting copula.

Theorem 2: Let $S_i \in \Omega$ $(i = 0, ..., n)$. The function Q defined in (7) provides a valid generator $\exp(-Q)$ for an Archimedean copula when,

- for $i = 1, ..., n$, $\theta_{0,i} = \sum_{k=1}^{i} \{\theta_{k-1} S_{k-1}(t_k) - \theta_k S_k(t_k)\}$, and
- when all the S_i are equal and the parameters θ_i $(i = 0, ..., n)$ are ordered as $1 \leq \theta_n \leq ... \leq \theta_0$.

The resulting copula has the following properties

- The same copula is generated whatever $\theta_{0,0}$. We can thus set $\theta_{0,0} = 0$, arbitrarily.
- $\exp(-Q)$ is a strict generator if and only if $\exp(-S_0)$ is strict.
- The resulting copula possesses $(n+1)$ parameters θ_i, $i = 0, ..., n$ supplementing those of the S_i. Each parameter θ_i quantifies locally the dependence in the $[t_i, t_{i+1})$ interval.
- Q is almost surely twice differentiable, except at the knots t_i, $i = 1, ..., n$, where it is continuous.
- Q is derivable at the knot t_i, if $\theta_{i-1} s_{i-1}(t_i) = \theta_i s_i(t_i)$, where $s_i(p) = \frac{dS_i(p)}{dp}$. When the same quantile function S is used across the intervals, this implies that the θ_i's must be equal.

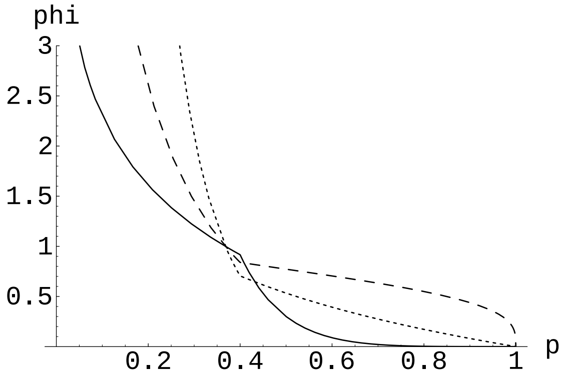

Figure 2. The generator $\phi = \exp(-Q)$ with Q defined in (7), $n = 1$ and $t_1 = 0.4$, is convex when $\theta_0 = 4$ and $\theta_1 = 1$ (solid line). It is not convex when (θ_0, θ_1) equals $(1, 4)$ (dotted line) or $(2, 0.3)$ (dashed line).

The condition on the intercepts insures the continuity of Q at the knots. As illustrated in Figure 2 with $S(p) = -\log\{-\log(p)\}$, the decreasing ordering condition $(1 \leq \theta_n \leq ... \leq \theta_0)$ is necessary to insure the convexity of the generator $\phi = \exp(-Q)$. When several $S_i(\cdot)$ functions are used in (7), the convexity condition is more complicated. Indeed, it is specific to the shape of the S_i functions that are involved.

The function Q defined in (7) only uses a single function S_i in the interval $[t_i, t_{i+1})$. Thus, its inverse Q^{-1} is simply

$$Q^{-1}(x) = \sum_{i=0}^{n} S_i^{-1}(\frac{x - \theta_{0,i}}{\theta_i}) I\{Q(t_i) \leq x < Q(t_{i+1})\}.$$

Then, the non-parametric bivariate Archimedean copula that uses Q as generator is, from Equation (6),

$$C(u, v) = \sum_{i=0}^{n} S_i^{-1}(\frac{x - \theta_{0,i}}{\theta_i}) I\{Q(t_i) \leq x < Q(t_{i+1})\}, \tag{8}$$

where

$$\begin{aligned} x &= -\log[\exp\{-\sum_{i=0}^{n} (\theta_{0,i} + \theta_i S_i(u)) I(t_i \leq u < t_{i+1})\} \\ &+ \exp\{-\sum_{i=0}^{n} (\theta_{0,i} + \theta_i S_i(v)) I(t_i \leq v < t_{i+1})\}]. \end{aligned}$$

C has a closed-form expression provided that all the S_i^{-1} $(i = 0, ..., n)$ have an analytical form.

4. Dependence properties

We now study dependence properties of the non-parametric Archimedean copula defined in Section 3. Using notations from Section 2.2, the distribution K of $C(U, V)$ becomes

$$K(p) = p + f_Q(p^+),$$

for $p \in [0, 1]$. When it exists, the density associated to K is

$$K'(p) = 1 + f'_Q(p).$$

K' is non-negative on $[0, 1]$, as $f'_Q(p)$ needs to be greater than or equal to -1 to fulfill the convexity condition introduced in (5).

Kendall's tau is then re-expressed as

$$\begin{aligned} \tau &= 1 - 4 \int_0^1 f_Q(p^+) dp \\ &= 1 - 4 \int_{\mathbb{R}} f^2(x) dx. \end{aligned}$$

Its value depends solely on the squared integral of the underlying univariate density.

When Q is the non-parametric generator defined in (7), tau becomes

$$\tau = 1 - 4 \sum_{i=0}^{n} \frac{1}{\theta_i} \int_{t_i}^{t_{i+1}} f_{S_i}(p^+) dp.$$

Since all $\theta_i > 1$, the larger the parameters, the larger τ will be. When the S_i are standardized quantile functions that do not contain any dependence parameter, the integrals $\int_{t_i}^{t_{i+1}} f_{S_i}(p^+) dp$ can be seen as weight factors w_i associated to θ_i in the sum. These weights are function of the considered function S_i and of the location of the knots t_i and t_{i+1}. The larger the size of the interval, the larger the weight. When the same function S is selected for all intervals, w_i solely depends on the location of the knots. It is then possible to select the knots so that all weights w_i are equal. In practice, we propose to use many equidistant knots so that w_i are all of the same very small order of magnitude.

One particular choice for S is the standardized log-log function $-\log(-\log(p))$. It corresponds to the generator $\phi(p) = -\log(p)$, which has been shown in Section 1 to generate independence. Plugged into (7) for all intervals, it creates a non-parametric copula which is Gumbel(θ_i) on each interval $[t_i, t_{i+1})$. With that choice, Kendall's tau becomes

$$\tau = \sum_{i=0}^{n} w_i \frac{\theta_i - 1}{\theta_i},$$

with $w_i = \{t_{i+1}^2(1 - 2\log t_{i+1}) - t_i^2(1 - 2\log t_i)\}$ where $\sum_{i=0}^{n} w_i = 1$. It is a weighted sum of locally-defined parameters

$$\tau_i = \frac{\theta_i - 1}{\theta_i}, \tag{9}$$

with weights w_i, $i = 0, ..., n$ depending only on the location of the knots. As shown in Table 1, for Gumbel's copula, the τ_i can be interpreted as Kendall's taus that are locally estimated based on values of $K(p)$ between t_i and t_{i+1}. With equidistant knots, the w_i have a bell-shaped curve having a maximum in the unit interval. The location of this maximum depends on the distance between knots in a non-algebraic way. Numerical procedures can also help to determine the series of $n+1$ knots that produces equal weights $w_i = \frac{1}{n+1}$. A local dependence interpretation for the parameters is thus possible.

5. Estimation procedure

Let (X_1, X_2) be a bivariate random vector for which m observations are available. Genest and Rivest (1993) showed that their joint empirical distribution at $(X_{1,k}, X_{2,k})$ is V_k $(k = 1, ..., m)$ where

$$V_k = \#\{(X_{1,j}, X_{2,j}) \leq (X_{1,k}, X_{2,k})\}/m. \tag{10}$$

Therefore, the empirical counterpart of K is the step function $K_m(V)$, such that $\forall V \in \mathbb{R}$,

$$K_m(V) = \#(V_j \leq V)/m.$$

Wang and Wells (2000) derived the corresponding estimator for bivariate censored data.

With the non-parametric quantile function (7), K becomes

$$K(p) = p + \sum_{i=0}^{n} \frac{1}{\theta_i} f_{S_i}(p^+) I(t_i \leq p < t_{i+1}).$$

Using a simple reparameterisation $\beta_i = 1/\theta_i$, parameters can be estimated based on the comparison of K_m with K across all m observed joint proportions

V_k, $k = 1, ..., m$. An L^2-norm comparison yields estimates

$$\widehat{\beta}_i = \arg\min[\sum_{k=1}^{m} \{K_m(V_k) - V_k - \sum_{i=0}^{n} \beta_i f_{S_i}(V_k^+) I(t_i \leq V_k < t_{i+1})\}^2], \qquad (11)$$
$$\text{with} \quad \beta_0 \leq ... \leq \beta_n \leq 1.$$

In matrix notations, we define the pseudo-observation m-vector $Y = \{K_m(V_k) - V_k, j = 1, ..., m\}$ and the $(m \times (n+1))$ design matrix $X = \{f_{S_i}(V_k^+) I(t_i \leq V_k < t_{i+1}), i = 0, ..., n, k = 1, ..., m\}$. The $(n+1)$ inequality constraints $\beta_0 \leq ... \leq \beta_n \leq 1$ can be expressed as $A^t \beta \geq 0$ for an appropriate choice of A. The estimation problem becomes a simple constrained least-squares estimation procedure

$$\arg_\beta \min (Y - X\beta)^t (Y - X\beta),$$

with

$$A^t \beta \geq 0.$$

The number and the location of the knots will influence the parameter estimates $\widehat{\beta}$. When $n = 1$, a single and constant dependence parameter is estimated, the least-squares estimator being an alternative to the moment-based estimator presented in Section 1. When knots are placed at all distinct V_k ($k = 1, ..., m$), this becomes a (constrained) interpolation problem. The level of smoothing of the fit increases as the number of knots decreases.

It also increases when the estimates of parameters corresponding to successive knots are constrained to be similar. Here, we propose to consider many equidistant knots and to control the smoothness of the fit by adding a penalty term to the least-squares problem. We follow Eilers and Marx (1996) and minimize the constrained *penalized* least squares

$$(Y - X\beta)^t (Y - X\beta) + \lambda \beta^t \Delta_r^t \Delta_r \beta$$
$$\text{subject to} \quad A^t \beta \geq 0, \qquad (12)$$

with $\Delta_1 \beta = (\beta_i - \beta_{i-1}, i = 1, ..., n)^t$ which builds up recursively to create $\Delta_r \beta = \Delta(\Delta_{r-1}\beta)$ that contains the r^{th} order finite differences between successive parameters. The weight λ determines the level of smoothing that is desired in the fit. A value $\lambda = 0$ corresponds to the constrained interpolation case with no smoothing. The correction $r = 1$ is a first-order correction as it tends to penalize any linear change in successive dependence parameters. When $\lambda \to +\infty$, a constant dependence parameter is estimated (one degree of freedom is left). With $r = 2$, quadratic changes are penalized, so that, when $\lambda \to +\infty$, only a linear change of dependence is possible over the joint distribution. The effective number of degrees of freedom is at most 2 in that case. It can actually be lower than 2, when constraints are activated. Although higher-order penalties are possible, taking $r = 2$ seems in general appropriate for most practical problems.

The optimization problem (12) can be solved using a quadratic programming algorithm with linear inequality constraints (Gill et al., 1981). We used the dual method of Goldfarb and Idnani (1983), implemented by Andreas Weingessel in the R language (Ihaka and Gentleman, 1996) to estimate parameters.

The effective number of degrees of freedom at the optimum $\widehat{\beta}$ depends on three factors: the number of knots n, the desired level of smoothing λ and the number of active constraints, i.e. those exhibiting an equality sign in (12). It can be assessed as follows. First, we examine the set of active constraints, and create the resulting equality constraint matrix $\widehat{A}$ by only keeping the corresponding columns in A. Thus one has $\widehat{A}^t\widehat{\beta} = 0$. Each active constraint induces a decrease of 1 in the number of degrees of freedom. Then, columns of the design matrix X corresponding to parameters affected by the active constraints are deleted. This is done using a QR decomposition of $\widehat{A}$ (Gill et al., 1981)

$$\widehat{A} = Q\begin{bmatrix} R \\ 0 \end{bmatrix} = [Q_R Q_N]\begin{bmatrix} R \\ 0 \end{bmatrix},$$

where Q is orthogonal and R is upper triangular. Q_N is a column basis for the null space of $\widehat{A}^t$: so that $Q_N^t\widehat{A} = 0$. Thus XQ_N is a reduced design matrix corresponding only to the unconstrained parameters. One then computes the influence (hat) matrix for the penalized least-squares problem using that design matrix as (Wood, 2000)

$$H = XQ_N(Q_N^t X^t X Q_N + \lambda Q_N^t \Delta_r^t \Delta_r Q_N)^{-1} Q_N^t X^t.$$

The part without the penalty λ in this expression is the well-known hat matrix for the ordinary least-squares problem. The other part is a negative quadratic penalty which value depends on the level of smoothing λ desired in the fit.

Finally, the *effective* number of degrees of freedom of the smoothing spline is estimated as

$$\mathrm{df} = \mathrm{trace}(H),$$

the sum of the diagonal elements of H. The coefficients df and λ are inversely related. The number of degrees of freedom provides an intuitive interpretation for the selected penalty λ. Observe that this number also encompasses the number and position of the knots as well as the number and position of the active constraints at the optimum. A strategy to find an optimal value for λ can be set up. This can be based on the minimization of the generalized cross-validation measure (Wahba, 1990)

$$GCV(\lambda) = \frac{1}{m} \sum_{k=1}^{m} \frac{(y_k - \widehat{y}_k))^2}{1 - \mathrm{df}/m}, \tag{13}$$

where $\widehat{y}_k$ are model predictions. In practice, since df depends on non-controllable factors (the active set at the optimum), it will only be possible to update the $GCV(\lambda)$ for several values of λ without re-estimating the parameters β when the active set is not modified. If the active set changes with λ, then parameters need to be re-estimated before computing the corresponding $GCV(\lambda)$ scores. The best λ value in a pre-specified set is then selected to minimize $GCV(\lambda)$.

5.1 *Variability estimation*

The constrained penalized least squares (PCLS) estimation algorithm does not tell us anything about the distributional properties of the parameter estimates. The availability of variance estimates for the parameter estimators and for the predictions could turn useful to assess the quality of the fit, and to evaluate the adequacy of any parametric proposal for the Archimedean copula. We estimate the distribution of the estimated parameters $\widehat{\beta}$ using the resampling-cases algorithm (Davison and Hinkley, 1999, p 264). Given the speed of the PCLS estimation algorithm, it is not tedious to use bootstrap methods to estimate variability of the parameter estimators. The procedure is as follows:

For a large number of replicates $r = 1, ..., R$,

1. sample $(X^*_{1,1}, X^*_{2,1}), ..., (X^*_{1,m}, X^*_{2,m})$ randomly with replacement from $\{(X_{1,k}, X_{2,k}), k = 1, ..., m\}$,
2. compute the empirical joint distribution of that sample $V^*_1, ..., V^*_m$ using (10),
3. apply the PCLS to the V^*_k ($k = 1, ..., m$) using the original knots and λ value, giving parameter estimates $\widehat{\beta}^*_r$.

When R is large, the covariance matrix and standard error of $\widehat{\beta}$ can be estimated using the corresponding estimates from the series of $\widehat{\beta}^*_r$, $r = 1, ..., R$. Non-parametric confidence limits for the predicted values $\widehat{Y}$ are also available using selected quantiles (2.5 and 97.5 %, usually to get 95% confidence bands) in the series $\{\widehat{Y}^*_r\}_{r=1}^{R}$.

An alternative and possibly more efficient bootstrap method is possible. It consists of sampling in step 1 directly from the predicted bivariate copula, whereas the empirical joint distribution was used in the initial proposal. Steps 2 and 3 are then unchanged.

6. Illustration in a lifetime study of Danish twins

The illustration is a large and non-censored cohort of 2872 pairs of Danish twins born 1870-1900. See Herskind et al. (1996) for more details on the data. Our interest is in the analysis of the dependence between the lifetimes of twins. A scatterplot of survival times is displayed in Figure 3. A few subjects die prior to 60 years of age. The death rate of individuals is rather low and uniform in the 20 to 60 years interval with most deaths occuring between 60 and 90 years old. A rank-based dependence measure, such as Kendall's tau (0.11) suggests an overall moderate positive dependence between the paired survival times. A visual inspection of Figure 3 reveals a more dense cloud of points on the main diagonal from 70-years old, onwards. This might indicate that the dependence could be larger at large survival times. Our formal non-parametric analysis of the dependence will confirm that observation.

Since our main interest is in the joint *survival* analysis of the twins, we fit the Archimedean copula models to the survival functions, instead of the distributions. To that end, we replace (10) with the joint empirical survival function

$$V_k = \#\{(X_{1,j}, X_{2,j}) \geq (X_{1,k}, X_{2,k})\}/m \tag{14}$$

and follow the same estimation procedure as proposed above with that function. The empirical joint survival distribution (14) counts 1527 distinct levels. Given that very large number, it is unnecessary to use the same number of knots. A well-selected reduced number of knots will lighten computations and have a negligible effect on the fit (Hastie et al., 2000, p. 163). Here, we used $n = 100$ knots, placed at every percent between 0 and 1.

We use the same quantile function $S(p) = -\log(-\log(p))$ of independence on every interval to create a non-parametric Archimedean copula along Eq. (7). Figure 4.a illustrates the fit of the 3 one-parameter Archimedean copula families of Table 1. In each case, the dependence parameter is estimated by equating the population version of Kendall's tau to its sample-based estimate. The 3 models fit differently to the data. The larger deviation to the data is seen in the middle of the plot (from 0.2 to 0.4). The fit of Clayton's family is better in the other parts of the plot, although it slightly underestimates the probability of low joint survival.

Figure 4.b presents a PCLS fit of the nonparametric copula. A second-order ($r = 2$) penalty was chosen, to constrain parameters β towards a linear decrease over increasing joint survival. The penalty in that fit ($\lambda = 10$) was chosen arbitrarily, leading to an effective number of degrees of freedom of 10.3. Indeed, the GCV estimate for that model ($2.7e-6$) was not much greater than the GCV ($1.6e-6$) of the model ($\lambda = 2e-4$; 47.2 df) that minimized (13). But the drop in

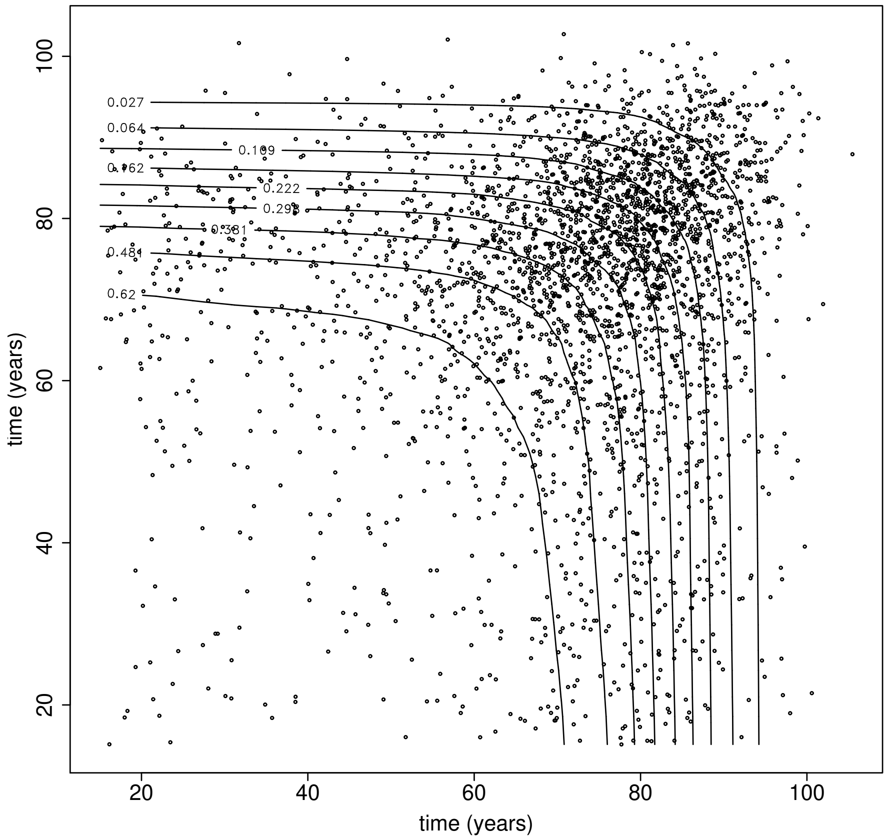

Figure 3. Joint survival time of pairs of Danish twins. Predicted contour of the joint survival distribution using non-parametric Archimedean copula with log-log quantile basis ($\lambda = 10$) is plotted as contour lines.

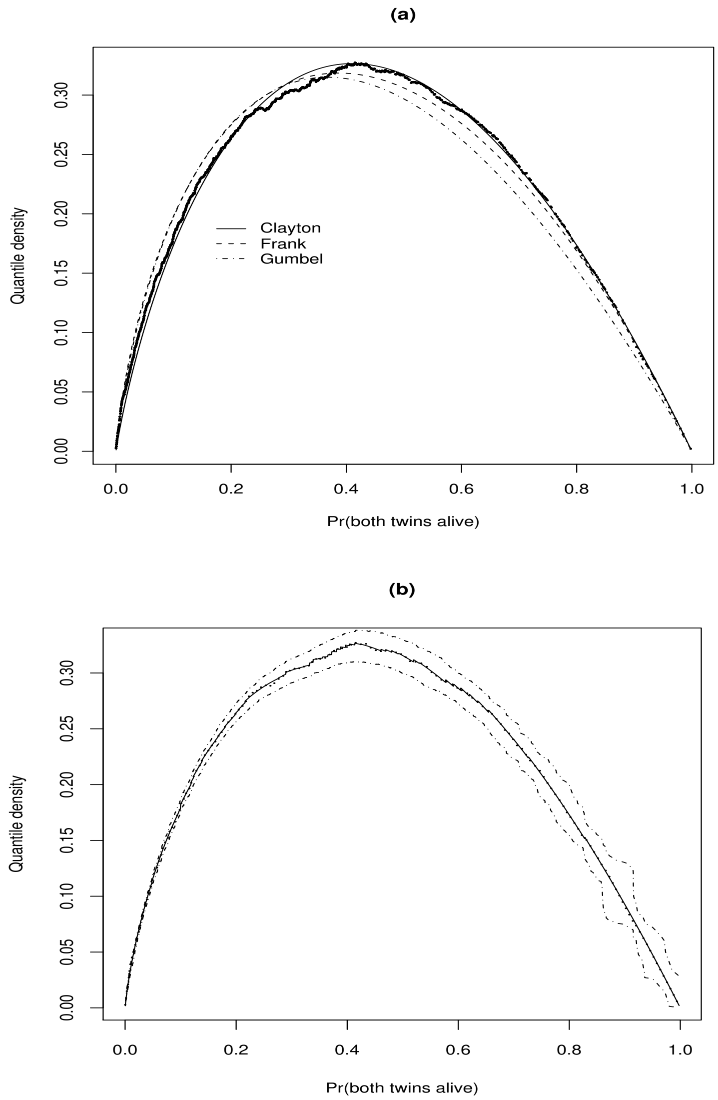

Figure 4. Empirical distribution $K_m(V_j)$ of the joint survival function V_j of Danish twins. a) Fit of selected 1-parameter Archimedean copulas with dependence parameter estimated from Kendall's tau. b) Constrained penalized ($\lambda = 10$) least-squares fit of a non-parametric Archimedean copula with log-log quantile basis (with 95% confidence bands).

number of degrees of freedom was big enough to be considered. The constrained model fits very well to the data. Indeed, it completely overlaps the empirical estimates in the plot. The predicted joint distribution from the non-parametric Archimedean copula is displayed as contour lines in Figure 3.

The non-smoothed copula model with $\lambda = 0$ (result not shown) counts 48 degrees of freedom; fifty-two out of the 100 parameters are captured in the set of activated constraints at the optimum. Hence, the constraints have an important effect on the smoothness of the curve, reducing already by 50% the number of effective parameters. This may explain why the GCV is minimized in a model that is almost not penalized (47.2 df). The minimal reduction in df at the optimum might also be due to the inadequacy of a second-order penalty for these data. However, the estimated parameters $\widehat{\beta}$ do not depart much from the linearity (result not shown) to expect a major change with a third-order penalty.

Confidence bands (95%) around the constrained model predictions were computed using the resampling-cases bootstrap method with $R = 500$. Bands are also in Figure 4.b. As expected, computation was very fast. Given the very large sample size ($N = 2872$), the standard error of all parameters was less that 5.8% of their mean and confidence bands were rather narrow. Despite this the fit of the non-parametric Archimedean copula was excellent. With this bootstrap method, we make no assumption about variance homogeneity and dependence structure of parameters. This proves useful to put the lights on the increasing variability of model predictions with increasing joint survival.

The local Kendall's tau estimates $\widehat{\tau_i}$ derived from (9) are in Figure 5. Despite its instability at low joint survival probabilities (ie, when relatively few pairs are still alive), the dependence decreases with increasing probabilities. The older the pair of twins, the greater the chance they will survive longer together. The fit of the non-parametric Archimedean copula with 10.3 df to the unconstrained smoother is very good. This good fit can motivate our selection of that "sub-optimal" penalty model. The few random discrepancies at low survivals give further arguments for preferring a model with fewer parameters than the optimal. Global Kendall's tau estimate (dotted line) is close to the mean of the local dependences, indicating that knots were well chosen to balance weights w_i.

Gumbel's copula, having a single parameter for the global dependence, does not fit well to the data. As shown in Figure 5, a constant dependence is assumed over the whole range of the joint distribution, which does not permit to put the increasing dependence over life time into perspective.

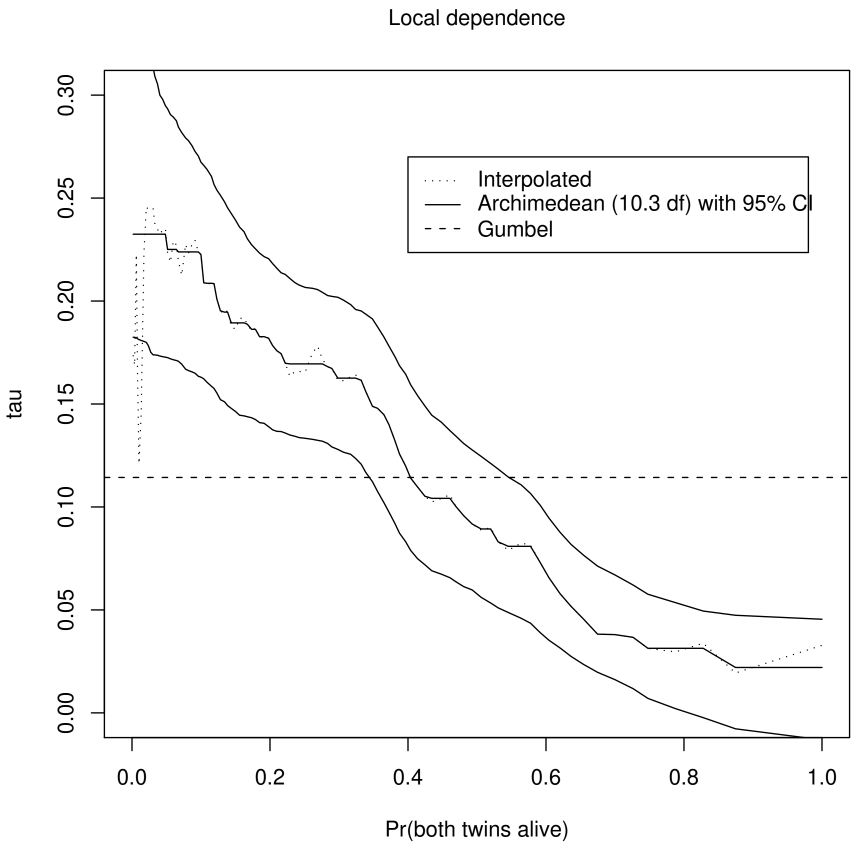

Figure 5. Local dependence as a function of the joint survival for Gumbel's copula, the unconstrained (interpolated) model and the non-parametric Archimedean copula counting 10.3 degrees of freedom.

7. Discussion

We have defined new bivariate non-parametric Archimedean copulas using local polynomials. We also provided a fast least-squares method to estimate parameters. We showed how the estimated parameters could be interpreted as local dependence measures when S was the log-log quantile. We provided an algorithm for smoothing that dependence over the joint distribution. The usefulness of the method was illustrated in a lifetime analysis of Danish twins. Other functions S than the log-log could be used as generator. See Section 2.1 for some examples. The dependence properties using alternative generators are yet to be established. Also, the selection of an optimal generator could be the topic of a future research.

The selected non-parametric copula is continuous but not derivable at a finite number of points being the knots. Also, dependence parameters have to be decreasing over increasing joint probabilities. An extension to non-linear quantile functions and to higher-order polynomials could be envisaged to overcome those limitations. It is also possible to estimate the generator ϕ directly using, for instance, penalized B-splines. However, interesting properties such as the possibility to have a closed-form expression for the joint distribution and the simple interpretation of parameters as local dependence measures might be lost in the operation.

Finally, we showed how to create univariate distributions from bivariate Archimedean copulas. The distributional properties of any particular family could be studied.

Acknowledgements

Data used for this research was provided by the Longitudinal Study of Aging in Danish Twins (LSADT), which is supported in part by funds from Duke University under an award from the U.S. National Institutes of Health, and by the Danish National Research Foundation. The findings, opinions and recommendations expressed therein are those of the authors and are not necessarily those of LSADT, Duke University, the National Institutes of Health or the Danish National Research Foundation.

Financial support from the contract 'Projet d'Actions de Recherche Concertées' nr 98/03-217 and from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged.

References

Davison, A. C. and Hinkley, D. V. (1999). *Bootstrap Methods and their Application.* Cambridge University Press.

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

Genest, C. and Mackay, R. J. (1986). Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *La revue canadienne de statistique* **14**, 145–159.

Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate archimedian copulas. *Journal of the American Statistical Association* **88**, 1034–1043.

Gilchrist, W. G. (2000). *Statistical Modelling with Quantile Functions.* Chapman and Hall/CRC.

Gill, P. E., Murray, W. and Wright, M. H. (1981). *Practical Optimization.* Academic Press.

Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming* **27**, 1–33.

Hastie, T., Tibshirani, R. and Friedman, J. (2000). *The Elements of Statistical Learning.* Springer Series in Statistics, Springer.

Herskind, A. M., McGue, M., Holm, N. V., Sorensen, T. I. A., Harvald, B. and Vaupel, J. W. (1996). The heritability of human longevity: a population-based study of 2872 danish twin pairs born 18701900. *Human Genetics* **97**, 319–323.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association* **53**, 814–861.

Nelsen, R. B. (1999). *An Introduction to Copulas.* Springer-Verlag.

Vandenhende, F. and Lambert, P. (2000). Modeling repeated ordered categorical data using copulas. *Discussion Paper* **00-25**, Institut de Statistique, Université Catholique de Louvain, Louvain–la–Neuve, Belgium. (ftp://www.stat.ucl.ac.be/pub/papers/dp/dp00/dp0025.ps).

Wahba, G. (1990). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics.

Wang, W. and Wells, M. T. (2000). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association* **95**, 62–76.

Wood, S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal Of the Royal Statistical Society, Series B* **62**, 413–428.