

Statistical Elucidation of Latent Structures via Copulas

Radu Craiu

Department of Statistical Sciences
University of Toronto

Latent Variable Models with Copulas

joint with Robert Zimmerman

Latent Variables (LV)

- ▶ The variable of interest W is sometimes impossible to measure directly
 - ▶ State of the economy
 - ▶ Traffic in a city
 - ▶ State of your health
 - ▶ State of a complex disease
- ▶ Instead, one measures
 - ▶ $\mathbf{Y} = (Y_1, \dots, Y_k)^T$ whose components are surrogates of W and each provide partial information about W
 - ▶ Covariate $\mathbf{X} \in \mathbb{R}^p$
- ▶ We are often interested in the explanatory power of $\mathbf{X}$ for W .

An Example with continuous outcomes

- ▶ The glomerular filtration rate (GFR) is the principal summary of kidney function, and it is not observed in routine care (Levey et al., 2020).
- ▶ Clinical practice uses filtration markers whose serum concentration rises as GFR falls.
- ▶ The surrogates for the GFR: serum cystatin C and serum β_2 -microglobulin (Shlipak et al., 2013; Inker et al., 2012, 2021).
- ▶ Data come from the 1999–2002 National Health and Nutrition Examination Survey (NHANES)

A bivariate continuous Example

- ▶ The sample is $n = 7468$ adults of median age 46 years (interquartile range 33–64), 52.2% of which are female.
- ▶ Covariates for W_i are age, sex, body mass index and systolic blood pressure, each standardised.
- ▶ Both markers are analysed on the reciprocal (clearance) scale, $q_i = 1/(\text{cystatin C})$ and $y_i = 1/(\beta_2\text{-microglobulin})$, in L/mg
- ▶ Kendall's τ between the two surrogates is 0.595.

The two surrogates

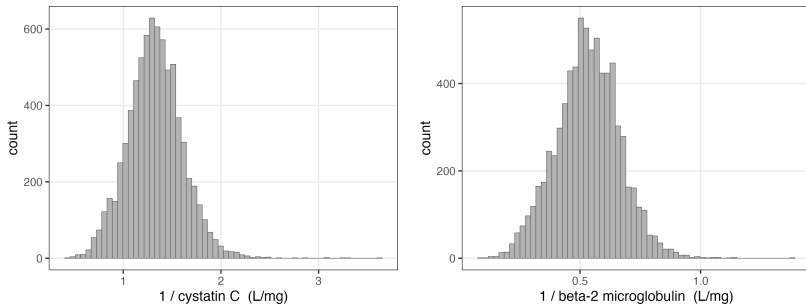


Figure: Clearance-scale histograms: $Q = 1/\text{cystatin C}$ (left) and $Y = 1/\beta_2\text{-microglobulin}$ (right).

Conditional independence LV model

- ▶ A canonical LV model:

$$Y_i \perp Q_i | W_i$$

$$Y_i | W_i \sim N(\mu_y + \lambda_y W_i, \sigma^2)$$

$$Q_i | W_i \sim \text{Gamma}(\alpha, \alpha \exp(-(\mu_q + \lambda_q W_i)))$$

$$W_i | \mathbf{x}_i \sim N(\mathbf{x}_i^\top \beta, \sigma_W^2)$$

- ▶ This implies that the two marginal regressions share a common random effect so they are marginally dependent (and conditionally independent)
- ▶ The induced dependence is not analytically available.

The two surrogates

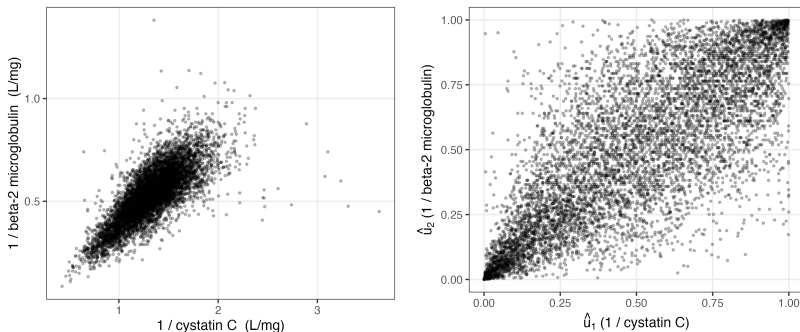


Figure: NHANES kidney surrogates on the clearance scale (left) and as pseudo-observations (right).

Empirical τ

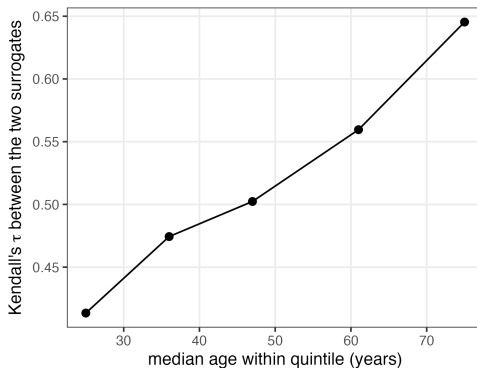


Figure: Kendall's τ between the two surrogates within quintiles of age.

Copulas: The Joys

- ▶ Copulas are mathematical devices used to model dependence between random variables regardless of their marginals.
- ▶ Copulas are useful for data fusion/integration as they lead to coherent joint models, even when the marginals are in different families or of different types.
- ▶ Copulas unlock information contained in the dependence part of the distribution (second-order) that complements the information in the marginals.
- ▶ Copulas extend statistical methods beyond the use of a multivariate Gaussian or Student.

At the root of it all, a theorem

- ▶ Copulas are distribution functions on $[0, 1]^d$ that model dependence between continuous random variables.
- ▶ **Sklar's Theorem:** If $Y_1, Y_2, \dots, Y_d$ are continuous r.v.'s with cdfs $F_1, \dots, F_d$, there exists an unique copula function $C : [0, 1]^d \rightarrow [0, 1]$ such that

$$H(t_1, \dots, t_d) = P(Y_1 \leq t_1, \dots, Y_d \leq t_d) = C(F_1(t_1), \dots, F_d(t_d)).$$

- ▶ The copula bridges the marginal distributions of $Y_1, \dots, Y_d$ with the joint distribution. It corresponds to a distribution on $[0, 1]^d$ with uniform margins.
- ▶ This can be extended to conditional distributions and copulas [Acar et al. \(2011\)](#); [Veraverbeke et al. \(2011\)](#); [Levi and Craiu \(2018\)](#)

$$P(Y_1 \leq t_1, \dots, Y_d \leq t_d | X) = C(F_1(t_1 | X), \dots, F_d(t_d | X) | X).$$

A conditional copula model

- $Q_i = 1/(\text{cystatin C})_i$, $Y_i = 1/(\beta_2\text{-microglobulin})_i$,
 $\mathbf{x}_i = (\text{age, female, BMI, SBP})_i$ standardised:

$$W_i \sim \mathcal{N}(\beta^\top \mathbf{x}_i, \sigma_W^2)$$

$$Q_i \mid W_i \sim \text{Gamma}(\alpha, \alpha e^{-(\mu_q + \lambda_q W_i)})$$

$$Y_i \mid W_i \sim \mathcal{N}(\mu_y + \lambda_y W_i, \sigma^2)$$

$$(Q_i, Y_i) \mid W_i \sim C\left(F^{\text{Gamma}}(\cdot \mid \alpha, \alpha e^{-(\mu_q + \lambda_q W_i)}), \Phi\left(\frac{\cdot - \mu_y - \lambda_y W_i}{\sigma}\right) \mid \theta(W_i, \xi)\right),$$

- Identifiability: $\|\beta\| = 1$ (so σ_W^2 is free) and $\lambda_y \geq 0$ (orients W as filtration).
- Consider a few families with different tail dependence and use a link/calibration function κ :

$$\theta(w, \xi) = \kappa(\xi_0 + \xi_1 w).$$

Conditional independence is a Copula LV

- ▶ The whole joint distribution of (Y, Q) is varying with W not just the marginals.
- ▶ The conditional copula captures the residual dependence on W after the marginal effects have been accounted for.
- ▶ The conditional independence model is obtained when using the independence copula.

Connection with factor copulas

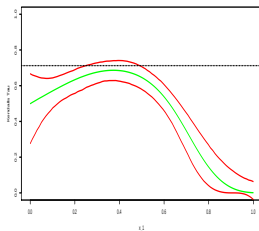
- ▶ Factor copula models ([Murray et al., 2013](#); [Krupskii and Joe, 2013, 2015](#); [Oh and Patton, 2017](#); [Nikolouloupoulos and Joe, 2015](#); [Choi et al., 2025](#)) also combine latent variables with copulas.
- ▶ In those models the observations are independent given the factors, and the factors are typically integrated out.
- ▶ Here the components of $\mathbf{Y}$ remain dependent given $\mathbf{W}$. The copula of that residual dependence, and the way its parameter varies with $\mathbf{W}$, is the object of interest.
- ▶ The latent variable is retained for inference.

A detour: Why the Conditional Copula?

- ▶ $Y_i|x \sim N(f_i(x), \sigma_i) \ x \in \mathbb{R}^2$
- ▶ True marginal means:
 - ▶ $f_1(x) = 0.6 \sin(5x_1) - 0.9 \sin(2x_2)$
 - ▶ $f_2(x) = 0.6 \sin(3x_1 + 5x_2)$
 - ▶ $\sigma_1 = \sigma_2 = 0.2, \textcolor{red}{X_1} \perp \textcolor{red}{X_2}.$
- ▶ Copula : $\theta(x) = 0.71$
- ▶ **Suppose x_2 is not observed so inference is based only on x_1**

A detour: Why the Conditional Copula?

- ▶ $Y_i|x \sim N(f_i(x), \sigma_i) \quad x \in \mathbb{R}^2$
- ▶ True marginal means:
 - ▶ $f_1(x) = 0.6 \sin(5x_1) - 0.9 \sin(2x_2)$
 - ▶ $f_2(x) = 0.6 \sin(3x_1 + 5x_2)$
 - ▶ $\sigma_1 = \sigma_2 = 0.2$, $X_1 \perp X_2$.
- ▶ Copula : $\theta(x) = 0.71$
- ▶ Suppose x_2 is not observed so inference is based only on x_1



- ▶ Additional details in Levi and Craiu, 2018.

A glance at computational difficulties

- ▶ W_i enters the likelihood only through $\lambda_q W_i$, $\lambda_y W_i$ and $\xi_1 W_i$. For any $C > 0$ the map

$$(\mathbf{W}, \beta, \lambda_q, \lambda_y, \xi_1, \sigma_W^2) \mapsto (C\mathbf{W}, C\beta, \lambda_q/C, \lambda_y/C, \xi_1/C, C^2\sigma_W^2)$$

leaves the law of $(Q_i, Y_i) \mid \mathbf{x}_i$ unchanged: the parameters are identified only up to this **orbit**.

- ▶ $\|\beta\| = 1$ picks one point on each orbit: β carries the direction of the covariate effect, the loadings its magnitude, σ_W^2 the variation in W beyond the covariates.
- ▶ The vanilla MH within Gibbs struggles with the dependence patterns in the posterior (e.g. between W , λ_q , λ_y and σ_W^2).

PX-DA helps with mixing

- ▶ Parameter expansion (Liu & Wu, 1999; Meng & van Dyk, 1999; Xu, Craiu, Sun & Paterson, 2016);
- ▶ Propose $\log(c) \sim \mathcal{N}(0, s_c^2)$ and set, holding β fixed,

$$(\mathbf{W}, \lambda_q, \lambda_y, \xi_1, \sigma_W^2) \mapsto (c\mathbf{W}, \lambda_q/c, \lambda_y/c, \xi_1/c, c^2\sigma_W^2).$$

- ▶ With s_c adapted to acceptance 0.44, the loadings and σ_W^2 mix like the other parameters.

An example with mixed responses

- ▶ Cardiotocography (CTG) is a medical procedure that monitors the fetal heart rate.
- ▶ The LV is the fetus' underlying state of health during birth, W .
- ▶ Our surrogate response is the bivariate vector (Q, Y) where
 - ▶ Q is the number of peaks (acceleration followed by a deceleration of heart beats) for the signal recorded by the CTG
 - ▶ Y is the log of mean short-term "beat-to-beat" variability (MSTV) where the short-term variability (STV) is obtained by measuring the time between successive R waves (cardiac systoles) of the fetus' electrocardiogram.
- ▶ The covariates are FM (fetal movement) and UC (uterine contraction), two continuous variables monitored during birth.

CTG: The LV Copula Model

- ▶ The copula model is specified, conditional on W_i ,

$$H(Y_i, Q_i | W_i) = C_{\theta_i}(F_Y(Y_i | W_i), F_Q(Q_i | W_i)), \quad \theta_i = \kappa(0.1(\xi_0 + \xi_1 W_i))$$

$$Y_i \sim N(\mu_y + \lambda_y W_i, \sigma^2); \quad Q_i \sim \text{Poisson}(\exp(\mu_q + \lambda_q W_i))$$

- ▶ $(Q_i, Y_i) | W_i$ has joint density

$$f_{(Q,Y)}(q, y | w) = f_y(y | w) \cdot [C_{q|y}(F_q(q | w), F_y(y | w)) - C_{q|y}(F_q(q - | w), F_y(y | w))],$$

where

$$C_{q|y}(u_q, u_y | w) = \frac{\partial}{\partial u_y} C(u_q, u_y | w).$$

CTG: The LV Copula Model

- ▶ Data Augmentation: Introduce latent variable Z such that

$$Q \stackrel{d}{=} F_q^{-1}(F_Z(Z)),$$

- ▶ Why: the copula of a mixed pair is not unique, and the density above depends on θ through a difference of conditional copulas.
- ▶ If (Z, Y) has copula C , integrating Z over the cell $(F_Z^{-1}(F_q(q-)), F_Z^{-1}(F_q(q))]$ returns the density above: the copula of (Q, Y) is preserved.
- ▶ Only $U = F_Z(Z)$ enters, so F_Z is a matter of convenience; with a Gaussian copula, $Z \sim N(0, 1)$ gives a truncated normal conditional.
- ▶ Craiu and Sabeti (2012); Smith and Khaled (2012).

The Augmented LV Copula Model for the CTG Example

- ▶ The augmented model for CTG data is

$$W_i \sim \mathcal{N}(\mathbf{x}_i^\top \beta, \sigma_W^2), \quad \|\beta\| = 1,$$

$$Z_i \sim \mathcal{N}(0, 1),$$

$$Y_i \mid W_i \sim \mathcal{N}(\mu_y + \lambda_y W_i, \sigma^2)$$

$$(Z_i, Y_i) \mid W_i \sim C^{\text{Gauss}} \left(\Phi(Z_i), \Phi \left(\frac{Y_i - \mu_y - \lambda_y W_i}{\sigma} \right) \mid \theta(W_i, \xi) \right)$$

$$P(Q_i = q \mid W_i, Z_i = z) = \mathbb{1}_{F_Z^{-1}(F_q(q - \mid \mu_q, \lambda_q, W_i)) < z \leq F_Z^{-1}(F_q(q \mid \mu_q, \lambda_q, W_i))}.$$

- ▶ Integrating Z_i out returns the mixed density of $(Q_i, Y_i) \mid W_i$: the copula of the observed pair is preserved.
- ▶ Same identification as in the continuous model.
- ▶ The same construction is used with Clayton, Frank and Gumbel copulas in place of C^{Gauss} .

CTG: The Augmented LV Copula Model

- ▶ Augmented posterior:

$$p(\boldsymbol{\eta}, \mathbf{W}, \mathbf{Z} \mid \text{data}) \propto \prod_{i=1}^n h(Z_i, y_i \mid W_i, \boldsymbol{\eta}) \\ \mathbb{1}_{a_i(W_i) < Z_i \leq b_i(W_i)} \phi(W_i - x_i^\top \boldsymbol{\beta}) \cdot p(\boldsymbol{\eta}),$$

with h the continuous density of $(Z_i, Y_i) \mid W_i$ and a_i, b_i the cell limits set by observation q_i and the Poisson margin.

- ▶ Conditioning everything on $\mathbf{Z}$ is simple but mixes badly.
- ▶ **Partially collapsed Gibbs** (Smith & Khaled 2012; van Dyk & Park 2008).

Sampling the augmented model

1. For (μ_q, λ_q) and $(\mu_y, \lambda_y, \log(\sigma)^2)$: adaptive RW Metropolis on the observed-data likelihood (Z integrated out)
3. $Z_i \mid \text{rest}$, by inversion: $V_i \sim \text{Unif}(C_{q|y}(F_q(q_i-), u_{y,i}), C_{q|y}(F_q(q_i), u_{y,i}))$,
 $Z_i = \Phi^{-1}(C_{q|y}^{-1}(V_i, u_{y,i}))$; closed form except Gumbel (bisection).
4. $\xi \mid \mathbf{Z}, \mathbf{W}$, rest: RW Metropolis on $\prod_i c(\Phi(Z_i), u_{y,i} \mid \theta(W_i, \xi)) p(\xi)$: a continuous copula likelihood, the payoff of the augmentation.
5. $\beta \mid \mathbf{W}, \sigma_W^2$: RW Metropolis on the unit sphere; then $\sigma_W^2 \mid \mathbf{W}, \beta \sim \text{InvGamma}$.
6. Each W_i : adaptive RW Metropolis, Z integrated out.
7. PX-DA group move $(\mathbf{W}, \lambda_q, \lambda_y, \xi_1, \sigma_W^2) \mapsto (c\mathbf{W}, \lambda_q/c, \lambda_y/c, \xi_1/c, c^2\sigma_W^2)$, Z integrated out.

Scoring the models: integrate out W

- ▶ Which density should a predictive criterion score? Two candidates for unit i with parameters η :
 - ▶ conditional on the latent variable: $p_{\mathcal{M}}(q_i, y_i \mid \eta, W_i)$;
 - ▶ **marginal over the latent variable:**

$$p_{\mathcal{M}}(q_i, y_i \mid \eta) = \int_{\mathbb{R}} p_{\mathcal{M}}(q_i, y_i \mid \eta, w) \frac{1}{\sigma_W} \phi\left(\frac{w - \beta^\top \mathbf{x}_i}{\sigma_W}\right) dw.$$

- ▶ The conditional version does not recover the correct copula family.
- ▶ The marginal version is the predictive density of a new subject with covariates $\mathbf{x}_i$ and is well calibrated for copula selection.
- ▶ Similar issues have been pointed out by [Celeux et al. \(2006\)](#).

Spotlight on dependence: A conditional WAIC

- Two conditional WAICs isolate the prediction of one component given the other (Levi and Craiu, 2018)

$$\text{CWAIC}_{y|q}(\mathcal{M}) = -2 \sum_{i=1}^n \log(\mathbb{E}[p_{\mathcal{M}}(y_i | q_i, \boldsymbol{\eta})]) + 2 \sum_{i=1}^n \text{Var}(\log(p_{\mathcal{M}}(y_i | q_i, \boldsymbol{\eta})))$$

$$\text{CWAIC}_{q|y}(\mathcal{M}) = -2 \sum_{i=1}^n \log(\mathbb{E}[p_{\mathcal{M}}(q_i | y_i, \boldsymbol{\eta})]) + 2 \sum_{i=1}^n \text{Var}(\log(p_{\mathcal{M}}(q_i | y_i, \boldsymbol{\eta})))$$

with $p_{\mathcal{M}}(q_i | y_i, \boldsymbol{\eta}) = p_{\mathcal{M}}(q_i, y_i | \boldsymbol{\eta}) / p_{\mathcal{M}}(y_i | \boldsymbol{\eta})$; in the mixed model $Y_i | \boldsymbol{\eta} \sim \mathcal{N}(\mu_y + \lambda_y \boldsymbol{\beta}^\top \mathbf{x}_i, \sigma^2 + \lambda_y^2)$ is closed form.

Spotlight on dependence: A conditional WAIC

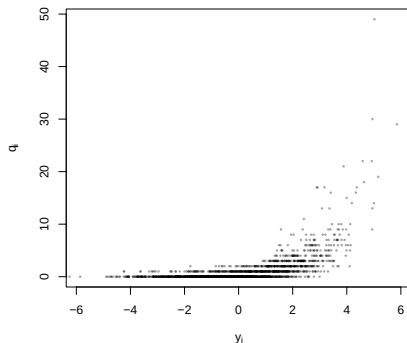
- ▶ $\frac{1}{2}(\text{CWAIC}_{q|y} + \text{CWAIC}_{y|q})$ is asymptotically equivalent to CCV for the marginal likelihood

$$\text{CCV}(\mathcal{M}) = \frac{1}{2} \left\{ \sum_{i=1}^n \log(\Pr(y_i | q_i, \mathcal{D}_{-i}, \mathcal{M})) + \sum_{i=1}^n \log(\Pr(q_i | y_i, \mathcal{D}_{-i}, \mathcal{M})) \right\}.$$

- ▶ For the kidney data we also report $\text{CVML} = \sum_i \log(p_{\mathcal{M}}(q_i, y_i | \mathcal{D}_{-i}))$.

Simulation: mixed responses, Gaussian copula

- ▶ $n = 2000$; $\mathbf{x}_i \stackrel{iid}{\sim} (\text{Unif}(-3, 1), \text{Unif}(-1, 1))$, standardised;
 $\beta = (1, 0.5)$, $\mu_q = -1$, $\lambda_q = 1$, $\mu_y = 0$, $\lambda_y = 1$, $\sigma^2 = 1$, $\xi = (3, 4)$.
- ▶ Conditional correlation $\tanh(0.1(3 + 4W_i))$ spans $(-0.97, 0.98)$; central 80% in $(-0.44, 0.79)$. Zero rate of the counts 0.62, max count 49.



Simulation Experiment: model selection

Five families fitted; two chains of 150,000 (burn-in 50,000). W -marginal criteria:

	Gaussian	Clayton	Frank	Gumbel	Independence
WAIC	10716.97	10753.75	10731.69	10738.09	10823.27
CWAIC _{$q y$}	3635.76	3672.42	3650.25	3656.41	3740.54
CWAIC _{$y q$}	6037.88	6074.88	6052.85	6058.59	6124.68

- ▶ All three criteria select the generating Gaussian copula: ahead of Frank by 15 units of WAIC, Gumbel by 21, Clayton by 37, independence by 106.
- ▶ The conditional criteria give the same ordering with the same gaps. PSRFs ≤ 1.02 .

Simulation Experiment I: posterior means

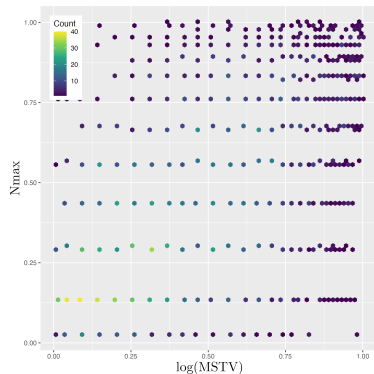
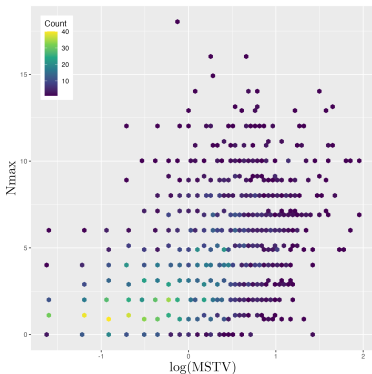
	β_1	β_2	μ_q	λ_q	μ_y	λ_y	σ^2	ξ_0	ξ_1	σ_W^2
Gaussian	0.89	0.45	-1.09	1.19	0.02	1.12	1.04	2.99	4.37	0.79
Clayton	0.89	0.45	-1.11	1.19	0.02	1.11	1.09	-9.37	13.37	0.79
Frank	0.89	0.46	-1.08	1.16	0.02	1.12	0.96	16.43	24.43	0.84
Gumbel	0.89	0.45	-1.09	1.20	0.02	1.10	1.12	-15.29	10.86	0.76
Independence	0.89	0.46	-1.11	1.06	0.02	1.14	0.33			1.27
True	1.0	0.5	-1.0	1.0	0.0	1.0	1.0	3.0	4.0	1.0

- The fit reports β on the unit sphere, relative sizes intact. The generating β has norm 1.12, so compare $\lambda_q, \lambda_y, \xi_1$ with 1.12, 1.12, 4.47 and σ_W^2 with 0.8.

What is identified, and what the convention adds

- ▶ The data inform the law of $(Q, Y) \mid \mathbf{x}$
 - ▶ the copula family and $\theta(w, \xi)$ at the quantiles of W (quantiles scale with W);
 - ▶ the covariate effects on each margin, $\lambda_q \beta$ and $\lambda_y \beta$: hence the direction of β , the relative sizes of β 's components, and λ_q / λ_y or $\frac{\lambda_q}{\lambda_q + \lambda_y}$;
 - ▶ $\mu_q, \mu_y, \sigma^2, \xi_0$;
 - ▶ the latent contributions per sd of W : $\lambda_q \sigma_W, \lambda_y \sigma_W, \xi_1 \sigma_W$; e.g. $\text{Var}(Y \mid \mathbf{x}) = \sigma^2 + \lambda_y^2 \sigma_W^2$.
- ▶ Signs are convention-free so $P(\lambda_q > 0 \mid \text{data})$ and whether an equal-tailed interval contains 0 do not depend on it

CTG: The data



$n = 2126$ cardiocograms (SisPorto); $q =$ number of peaks ($N_{\max}$), $y = \log$ mean short-term variability. Left: original scales. Right: pseudo-observations rank/ $(n + 1)$, the copula scale; the count margin produces the horizontal bands. Marginal Kendall's $\tau = 0.41$: positive, while the conditional copula given W will turn out negative.

CTG: Estimates

- ▶ Compare Gaussian, Clayton, Frank, Gumbel and independence copulas
- ▶ All criteria select the Gaussian copula
- ▶ Posterior means (95% intervals) under the Gaussian copula, $\|\beta\| = 1$

	β_1 (FM)	β_2 (UC)	μ_q	λ_q	μ_y	λ_y	σ^2	ξ_0	ξ_1	σ_W^2
Mean	0.403	0.914	1.267	0.204	0.073	0.249	0.038	-7.53	-1.04	5.92
2.5%	0.307	0.869	1.234	0.180	0.046	0.221	0.022	-12.65	-1.99	4.72
97.5%	0.495	0.952	1.300	0.229	0.100	0.276	0.060	-4.18	-0.29	7.44

- ▶ $\theta(w, \xi) = \tanh(0.1(-7.5 - 1.0w))$, $\text{sd}(W) \approx 2.6$: negative association between Nmax and STV throughout; $\theta = -0.64$ at the median W , -0.79 at the 95% quantile, weakening to -0.25 (interval reaching 0) at the 5% quantile $W \approx -4.7$. Both loadings positive, so low W means few peaks and low STV: poorer fetal state. A reading: in healthy state the alternation of quiet and active states moves peaks and variability in opposite directions.

Kidney data: model comparison

Seven families, two chains of 200,000 (burn-in 100,000)

copula	WAIC	Δ WAIC	CVML
Frank	-17814.0	0.0	8907.0
Gaussian	-17656.3	157.7	8828.2
Gumbel	-17589.0	225.1	8794.5
Survival Clayton	-17577.8	236.2	8788.9
Joe	-17536.1	277.9	8768.1
Clayton	-17195.0	619.0	8597.5
Independence	-17075.0	739.0	8537.5

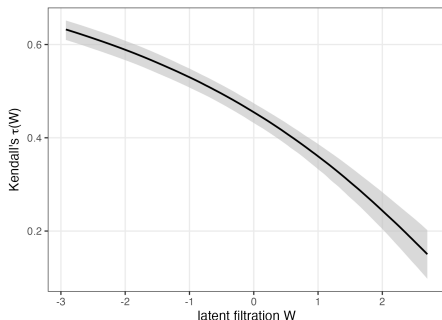
- Every copula improves on independence

Kidney data: the selected Frank model

parameter	mean	95% c.i.
λ_q	0.124	(0.119, 0.128)
λ_y	0.073	(0.071, 0.076)
α	40.7	(38.1, 43.6)
σ	0.093	(0.089, 0.096)
ξ_0	49.6	(46.2, 52.9)
ξ_1	-13.3	(-15.0, -11.6)
β_{age}	-0.961	(-0.970, -0.952)
β_{female}	0.177	(0.146, 0.208)
β_{BMI}	-0.208	(-0.237, -0.179)
β_{SBP}	-0.019	(-0.057, 0.018)
σ_W^2	0.580	(0.486, 0.684)

- Both loadings positive: filtration is represented by W . Posterior means of W_i have Spearman correlation 0.75 with creatinine-based eGFR (not used in the fit); age has a negative effect on W (intuitive).
- Age is the strongest covariate, then BMI; women filter more than men of the same age and build; SBP interval covers zero.
- $\sigma_W^2 = 0.58$ and $\text{Var}(\beta^T X) = 1.04$ so the four covariates account for about 64% of the variation in filtration.

Dependence strengthens as filtration falls



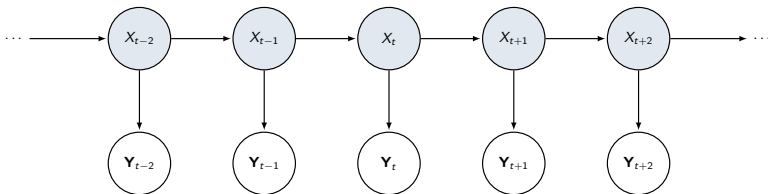
- ▶ Kendall's τ of the conditional copula falls from about 0.60 at the 5th percentile of fitted W to about 0.25 at the 95th.
- ▶ $\Pr(\xi_1 > 0 \mid \text{data})$ is numerically zero under every family except Clayton (0.03).
- ▶ Combining the two markers adds least information exactly where the assessment matters most.

Copulas for serially correlated data with hidden structures

Joint with Robert Zimmerman and Vianey Leos Barajas

Hidden Markov Models: A Primer

- ▶ A hidden Markov model (HMM) pairs an observed time series $\{\mathbf{Y}_t\}_{t \geq 1} \subseteq \mathbb{R}^d$ with a Markov chain $\{X_t\}_{t \geq 1}$ on some state space $\mathcal{X}$, such that the distribution of $\mathbf{Y}_s \mid X_s$ is independent of $\mathbf{Y}_t \mid X_t$ for $s \neq t$:



- ▶ $\mathbf{Y}_{t,h} \mid \{X_t = k\} \sim f_{k,h}(\cdot \mid \lambda_{k,h}) \quad \forall h = 1, \dots, d$
- ▶ $\{X_t\}$ is a Markov process (finite state space $\mathcal{X}$) with initial probability mass distribution $\{\pi_i\}_{i \in \mathcal{X}}$ and transition probabilities $\{\gamma_{i,j}\}_{i,j \in \mathcal{X}}$

Inferential aims for HMMs

- ▶ Typically, the chain $\{X_t\}_{t \geq 1}$ is partially or completely unobserved.
- ▶ The hidden states can correspond to a precise variable (occupancy data) or might be postulated (psychology, ecology, etc)
- ▶ **Aim 1:** Model the data generating mechanism (Nasri et al., 2020)
- ▶ **Aim 2:** Decode (i.e., classify) or predict the X_t 's from the observed data.

Examples

- ▶ A tri-axial accelerometer captures a shark's acceleration with respect to three positional axes depending on the shark's activity (resting, hunting, attacking).
- ▶ Hospitals repeatedly record a patient's vitals post-surgery to understand the state of their recovery (stable, complications, complete).
- ▶ In-game team statistics like shots on goal, forward or back passes in a soccer football match are changing with the "momentum" of the team (defensive, offensive, passive).

Fusion of Multiple Data Sources

- ▶ In the real-world applications above, various sensors capture multiple streams of data, which are “fused” into a multivariate time series $\{\mathbf{Y}_t\}_{t \geq 1}$.
- ▶ In such situations, the components of any $\mathbf{Y}_t = (Y_{t,1}, \dots, Y_{t,d})$ cannot be assumed independent (even conditional on X_t).
- ▶ The corresponding assumption for HMMs – that of contemporaneous conditional independence (Zucchini et al, 2017) – is often violated.
- ▶ Instead, it is common to assume that $\mathbf{Y}_t$ follows a multivariate Gaussian distribution, but this places limits on marginals and dependence structures.
- ▶ What if the strength/type of dependence between the components of $\mathbf{Y}_t$ could be informative about the underlying state X_t ?

Occupancy Data

- ▶ The ability to detect whether a room is occupied using sensor data (such as temperature and CO₂ levels)
- ▶ Consider three publicly-available labelled datasets presented by Candanedo and Feldheim (2016) which contain multivariate time series of four environmental measurements (light, temperature, humidity, CO₂) and one derived metric (the humidity ratio).

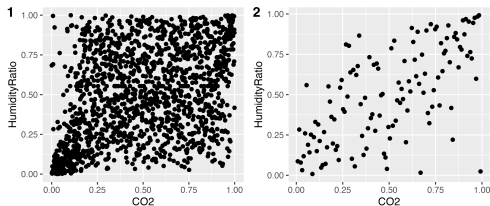


Figure: Pseudo-observations (after marginals are fitted) computed from unoccupied (Panel 1) and occupied (Panel 2) subsets.

Copulas Within HMMs

- ▶ Our model consists of an HMM $\{(\mathbf{Y}_t, X_t)\}_{t \geq 1} \subseteq \mathbb{R}^d \times \mathcal{X}$ in which the state-dependent distributions are copulas:

$$\mathbf{Y}_t \mid (X_t = k) \sim H_k(\cdot) = \underbrace{C_k\left(F_{k,1}(\cdot; \lambda_{k,1}), \dots, F_{k,d}(\cdot; \lambda_{k,d})\right)}_{\text{depends on the hidden state value } k} \mid \theta_k.$$

- ▶ $C_k(\cdot, \dots, \cdot \mid \theta_k)$ is a d -dimensional parametric copula
- ▶ $\{X_t\}_{t \geq 1}$ is a Markov process on finite state space $\mathcal{X} = \{1, 2, \dots, K\}$ and K is known.
- ▶ In this model, virtually all aspects of the state-dependent distributions are allowed to vary between states

Estimation with missing data

- ▶ Data consist in observed $\mathbf{Y}_{1:T}$ and missing $X_{1:T}$
- ▶ Parameters are $\eta = \{\lambda_{h,k}\}_{h=1:d}^{k=1:T} \cup \{\theta_k\}_{k=1:T} \cup \{\gamma_{i,j}\}_{i=1:K}^{j=1:K} \cup \{\pi_j\}_{j=1:K}$.
- ▶ The complete-data log-likelihood for one trajectory of the copula HMM is given by

$$\begin{aligned} \ell_{\text{com}}(\eta \mid \mathbf{y}_{1:T}, X_{1:T}) = & \pi_{X_1} + \sum_{t=2}^T \log \gamma_{X_{t-1}, X_t} + \sum_{h=1}^d \log f_{X_t, h}(y_{t, h}; \lambda_{X_t, h}) \\ & + \sum_{t=1}^T \log c_{X_t}(F_{X_t, 1}(y_{t, 1}; \lambda_{X_t, 1}), \dots, F_{X_t, 1}(y_{t, d}; \lambda_{X_t, d}) \mid \theta_{X_t}). \end{aligned} \quad (1)$$

Inference for HMMs Via the EM Algorithm

- ▶ Without copula, the estimation is done via the EM algorithm (aka Baum-Welch)

E-step Compute $Q(\eta|\eta^{(s)}) = E[l_{com}(\eta|\mathbf{Y}_{1:T}, X_{1:T})|\eta^{(s)}, \mathbf{Y}_{1:T}]$

M-step Set $\eta^{(s+1)} = \arg \max_{\eta} Q(\eta|\eta^{(s)})$

- ▶ The complete-data log-likelihood is written in terms of the state membership indicators $U_{k,t} = \mathbb{1}_{X_t=k}$ and $V_{j,k,t} = \mathbb{1}_{X_{t-1}=j, X_t=k}$
- ▶ In the **E-Step**, these indicators are estimated by the conditional probabilities $\hat{u}_{k,t} = \mathbb{P}(X_t = k | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$ and $\hat{v}_{j,k,t} = \mathbb{P}(X_{t-1} = j, X_t = k | \mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$, which are computed based on current parameter estimates
- ▶ This only requires evaluating the state-dependent densities at each of the observations $\mathbf{y}_1, \dots, \mathbf{y}_T$ (this is “OK”)

A glance at computational challenges

- ▶ In the **M-Step**, the resulting complete-data log-likelihood is maximized with respect to all parameters in the model simultaneously
 - ▶ Only for the simplest univariate models do the state-dependent MLEs exist in closed form; otherwise, one must resort to numerical methods (**this is hard and unstable!**)
 - ▶ Evaluating a copula density $c_k(\cdot, \dots, \cdot \mid \theta_k)$ in high dimensions is slow
 - ▶ When the state-dependent distributions in an HMM are copulas, performing the M-Step directly requires the evaluation of

$$\operatorname{argmax}_{\{\theta_k\}, \{\lambda_{k,h}\}} \left\{ \sum_{k=1}^K \sum_{t=1}^T \hat{u}_{k,t} \left[\log c_k \left(F_{k,1}(y_{t,1}; \lambda_{k,1}), \dots, F_{k,d}(y_{t,d}; \lambda_{k,d}) \mid \theta_k \right) + \sum_{h=1}^d \log f_{k,h}(y_{t,h}; \lambda_{k,h}) \right] \right\}$$

- ▶ This is very unstable (and slow)

Inference Functions for Margins

- ▶ Likelihood-based inference for copulas is easier when the goal is to estimate θ alone in the presence of known margins
- ▶ Why not perform inference on the marginal distributions first, and then on the copula itself?
- ▶ In the context of iid data, this is exactly the **inference functions for margins** (IFM) approach of Joe and Xu (1996):
 - ▶ First estimate each λ_h by its “marginal MLE” $\hat{\lambda}_h$ given $\{Y_{t,h}\}_{t \geq 1}$, for $h \in \{1, \dots, d\}$
 - ▶ Then estimate θ assuming fixed marginals $F_1(\cdot; \hat{\lambda}_1), \dots, F_d(\cdot; \hat{\lambda}_d)$
- ▶ One can show that the IFM estimator is consistent and asymptotically normal (although relatively less efficient than the MLE)

IFM Step

- For each $j \in \mathcal{X}$, estimate the initial distribution and transition probabilities:

$$\delta^{(s+1)} = (\hat{u}_{1,1}^{(s)}, \dots, \hat{u}_{K,1}^{(s)})$$

and

$$\gamma_{j,\cdot}^{(s+1)} = \left(\frac{\sum_{t=2}^T \hat{v}_{j,1,t}^{(s)}}{\sum_{k=1}^K \sum_{t=2}^T \hat{v}_{j,k,t}^{(s)}}, \dots, \frac{\sum_{t=2}^T \hat{v}_{j,K,t}^{(s)}}{\sum_{k=1}^K \sum_{t=2}^T \hat{v}_{j,k,t}^{(s)}} \right).$$

- For each $k \in \{1, \dots, K\}$ and $h \in \{1, \dots, d\}$, estimate the marginal parameters

$$\lambda_{k,h}^{(s+1)} = \arg \sup_{\lambda} \sum_{t=1}^T \hat{u}_{k,t}^{(s+1)} \cdot \log(f_{k,h}(y_{t,h}; \lambda)).$$

- For each $k \in \{1, \dots, K\}$, estimate the copula parameters

$$\tilde{\theta}_k^{(s+1)} = \arg \sup_{\theta} \sum_{t=1}^T \hat{u}_{k,t}^{(s+1)} \cdot \log \left(c_k \left(F_{k,1}(y_{t,1}; \lambda_{k,1}^{(s+1)}), \dots, F_{k,d}(y_{t,d}; \lambda_{k,d}^{(s+1)}) \mid \theta \right) \right).$$

New problems

- ▶ The EIFM algorithm is not an GEM algorithm

$$\sum_{t=1}^T \hat{u}_t \cdot \log \left(f_h(y_{t,h}; \lambda_h^{(s)}) \right) \leq \sum_{t=1}^T \hat{u}_t \cdot \log \left(f_h(y_{t,h}; \lambda_h^{(s+1)}) \right), \quad h \in \{1, \dots, d\} \quad (2)$$

does not imply

$$\begin{aligned} & \sum_{t=1}^T \hat{u}_t \cdot \log \left(c \left(F_1(y_{t,1}; \lambda_1^{(s)}), \dots, F_d(y_{t,d}; \lambda_d^{(s)}) \mid \theta^{(s)} \right) \right) \\ & \leq \sum_{t=1}^T \hat{u}_t \cdot \log \left(c \left(F_1(y_{t,1}; \lambda_1^{(s+1)}), \dots, F_d(y_{t,d}; \lambda_d^{(s+1)}) \mid \theta^{(s)} \right) \right). \end{aligned}$$

- ▶ The EIFM algorithm will converge (to a local or global maximum).
- ▶ The estimator is consistent and asymptotically normal (under mild regularity conditions).
- ▶ EIFM as a version of the ES algorithm of Elashoff and Ryan (2004).
- ▶ Use asymptotic theory of M-estimators for HMMs Jensen (2011).

Implementation of EIFM

- ▶ K is assumed known
- ▶ An initial clustering algorithm may be used in which the observed multivariate data follow vine copulas (Sahin and Czado, 2022).
- ▶ We consider the k -means algorithm for clustering.
- ▶ Initial parameter values for the copula(s) are obtained using a Gaussian copula
- ▶ If marginals are Gaussian this means fitting a multivariate normal for each cluster.

Does This Work?

- ▶ For $T \in \{100, 1000, 5000\}$ and $d \in \{2, 5, 10\}$, we simulated a d -dimensional time series of length T from the 2-state HMM

$$\mathbf{Y}_t \mid (X_t = 1) \sim C_{\text{Frank}} \left((\mathcal{N}(\mu_{1,h} = -h, 1))_{h=1}^d \mid \theta_1 = 3 \right)$$

$$\mathbf{Y}_t \mid (X_t = 2) \sim C_{\text{Clayton}} \left((\mathcal{N}(\mu_{2,h} = h, 1))_{h=1}^d \mid \theta_2 = 3 \right)$$

and estimated $\boldsymbol{\eta} = (\mu_{1,1}, \dots, \mu_{2,d}, \theta_1, \theta_2)$ using both approaches

- ▶ Applied to the basic EM algorithm, R's `optim` with L-BFGS-B (i.e., quasi-Newton with box constraints) typically fails as soon as $d \geq 3$
 - ▶ The procedure is extremely sensitive to initial values and requires $\hat{\boldsymbol{\eta}}^{(0)} \approx \boldsymbol{\eta}$ just to avoid overflow
 - ▶ This kind of tuning is very tedious or impossible in high dimensions

Does This Work?

- We keep track of the **time** (in seconds) until the algorithm converges, and the **L_2 error** of the resulting estimate, $\epsilon = \|\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}\|_2$
 - We used the `lbfgsb3c` package, which is more stable than `optim`

	$d = 2$	$d = 5$	$d = 10$
$T = 100$	111.9 s, $\epsilon = 0.14$	123.4 s, $\epsilon = 299.98$	111.8 s, $\epsilon > 10^9$
$T = 1000$	166.6 s, $\epsilon = 0.63$	169.5 s, $\epsilon > 10^{11}$	418.23 s, $\epsilon = 725.06$
$T = 5000$	?	?	?

Table: EM Algorithm

	$d = 2$	$d = 5$	$d = 10$
$T = 100$	5.1 s, $\epsilon = 2.9$	3.0 s, $\epsilon = 0.94$	4.2 s, $\epsilon = 0.58$
$T = 1000$	34.4 s, $\epsilon = 0.57$	22.9 s, $\epsilon = 0.60$	34.4 s, $\epsilon = 0.80$
$T = 5000$	172.6 s, $\epsilon = 0.13$	106.2 s, $\epsilon = 0.12$	168.7 s, $\epsilon = 0.19$

Table: EFM Algorithm

Numerical Experiment I

► Generative model:

$$\mathbf{Y}_i \mid (X_i = k) \sim C_k \left(SN(\cdot; \xi_{k,1}, \omega_{k,1}, \alpha_{k,1}), SN(\cdot; \xi_{k,2}, \omega_{k,2}, \alpha_{k,2}) \mid \tau_k \right),$$

for $k \in \{1, \dots, 4\}$.

State	Copula family	τ_k	$\xi_{k,1}$	$\omega_{k,1}$	$\alpha_{k,1}$	$\xi_{k,2}$	$\omega_{k,2}$	$\alpha_{k,2}$
1	Clayton	0.2	-4	1	5	-1	1	-3
2	B4	0.4	-2	1	3	2	1	-3
3	Gaussian	0.6	0	1	5	3	1	-5
4	$t_{(\nu=5)}$	0.8	2	1	3	4	1	-5

Table: True parameters for the state-dependent distributions.

Numerical Experiment I

T		500	1000	2500	5000
Stopping Rule Tolerance:	0.01	14	24	23	15
	0.001	17	26	25	17
	0.0001	36	59	62	39
	0.00001	230	115	460	269
Classifier:	k -means	0.9020	0.9090	0.9200	0.9196
	Local decoding	0.9640	0.9640	0.9696	0.9732

For each $T \in \{500, 1000, 2500, 5000\}$: (Top rows) Number of iterations taken by the EIFM algorithm applied to $\{Y_t\}_{1 \leq t \leq T}$ before stopping using L_1 -norm tolerances in $\{0.01, 0.001, 0.0001, 0.00001\}$. (Bottom rows) Classification accuracy of initial k -means clustering and local decoding with parameter estimates provided by the EIFM algorithm.

Numerical Experiment II

- ▶ The state-dependent distributions are Markov trees in which all conditional relationships are independent.



Figure: Markov trees for state 1 (left) and state 2 (right)

- ▶ The state-dependent distributions have densities supported on $\mathbb{R}^5$

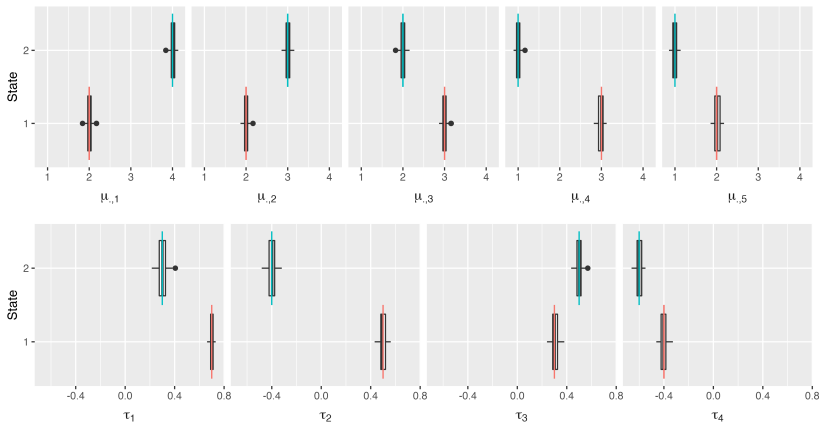


Figure: Parameter estimates based on 100 independent simulations and EIFM algorithm runs for the 5-dimensional 2-state HMMs

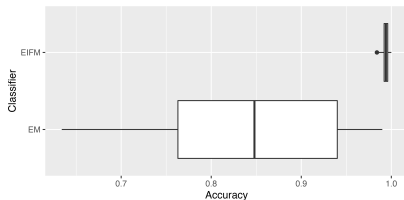


Figure: Accuracy across repetitions using initial independence model (bottom) and copula model (top)

Occupancy Data

Classifier	Train	Test 1
k-means clustering	0.865	0.818
Independence copulas within HMM	0.895	0.846
BB7/Tawn copulas within HMM	0.900	0.852

The Past & Future

- ▶ Copulas offer a way to bypass the paucity of available joint distributions.
- ▶ Copulas allow the integration of multiple (dependent) sources of information/data via joint modeling
- ▶ Joint models can be used for prediction/imputation of an expensive variable given values for cheaper ones.
- ▶ So far have been used to further empower multivariate regressions, time series, HMMs, LV models, etc
- ▶ Computational challenges, especially in higher dimensions
- ▶ Papers available here: raducraiu.com/publications/

References

- ACAR, E. F., CRAIU, R. V. and YAO, F. (2011). Dependence Calibration in Conditional Copulas: A Nonparametric Approach. *Biometrics* **67** 445–453.
- CELEUX, G., FORBES, F., ROBERT, C. P. and TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models .
- CHOI, J., CHUNG, H. C., GAYNAOVA, I. and NI, Y. (2025). Bayesian segmented gaussian copula factor model for single-cell sequencing data. *Bayesian Analysis* **1** 1–26.
- KRUPSKII, P. and JOE, H. (2013). Factor copula models for multivariate data. *Journal of Multivariate Analysis* **120** 85–101.
- KRUPSKII, P. and JOE, H. (2015). Structured factor copula models: Theory, inference and computation. *Journal of Multivariate Analysis* **138** 53–73.
- LEVI, E. and CRAIU, R. V. (2018). Bayesian inference for conditional copulas using Gaussian Process single index models. *Computational Statistics & Data Analysis* **122** 115–134.
- MURRAY, J. S., DUNSON, D. B., CARIN, L. and LUCAS, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association* **108** 656–665.

References (cont.)

- NIKOLOULOPOULOS, A. K. and JOE, H. (2015). Factor copula models for item response data. *Psychometrika* **80** 126–150.
- OH, D. H. and PATTON, A. J. (2017). Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics* **35** 139–154.
- VERAVERBEKE, N., OMELKA, M. and GIJBELS, I. (2011). Estimation of a Conditional Copula and Association Measures. *Scandinavian Journal of Statistics* **38** 766–780.