Bayesian Inference for Conditional Copula models

Radu Craiu

Department of Statistical Sciences
University of Toronto

Joint with Evgeny Levi (Toronto), Avideh Sabeti (Toronto) and Mian Wei (Toronto)

CRM, Montreal
January 27, 2017

Outline

Copulas

- Copula functions are used to model dependence between continuous random variables.
- (Sklar,'59) If Y_1, Y_2 are continuous r.v.'s with distribution functions (df) F, G, there exists an unique copula function $C : [0,1] \times [0,1] \to [0,1]$ such that

$$H(t,s) = \Pr(Y_1 \leq t, Y_2 \leq s) = C(F(t), G(s)).$$

- C is a distribution function on $[0,1]^2$ with uniform margins.
- The copula bridges the marginal distributions with the joint distribution.

Inference for Copula Models

- Early on: abundance of theoretical developments: construction of new copula families and connections with dependence concepts (NA, PQD/NQD, PRD/NRD, etc). Joe ('97), Nelsen ('06).
- Statistical inference for constant copula models:
  - Joint Maximum Likelihood: numerical methods
  - Two-stage approach: Joe (JMVA, '05)
  - Semiparametric approach: Genest, Khoudi & Rivest (Bmka, '95)
  - Nonparametric approach: Pickands (BISI,'81); Capéraà, Fougères & Genest (Biomka, '97) .
- Copula goodness-of-fit and selection (Genest, Remillard & Beaudoin, IME '09; Genest, Quessy & Remillard, SJS '06; Fermanian, JMVA '05; Berg, Eur. J. of Finance, '09).

Conditional Copula

- Consider a random sample $\{x_i \in \mathbf{R}^d, y_{1i} \in \mathbf{R}, y_{2i} \in \mathbf{R}\}_{1 \leq i \leq n}$ and suppose $F_X(y_1)$ and $G_X(y_2)$ are the unknown marginal conditional cdf's.
- The bivariate conditional copula (CC) of $(Y_1, Y_2)|X = x$, is the conditional joint distribution function of $U = F_x(Y_1)$ and $V = G_x(Y_2)$ given $X = x$ (Patton, Int'l Econ. Rev. '06)

$$H_x(t, s) = C_x(F_x(t), G_x(s))$$

- The parametric bivariate CC model assumes there is a parametric family $\mathcal{C} = \{C_\theta : \ \theta \in \Theta\}$ s.t.

$$C_x(F_x(t), G_x(s)) = C_{\theta(x)}(F_x(Y_1), G_x(Y_2)).$$

- The simplifying assumption:

$$C_x(F_x(y_1), G_x(y_2)) = C(F_x(y_1), G_x(y_2)).$$

Why CC? Understanding the Dependence Pattern

- We are interested in understanding the covariate effect on the dependence pattern between responses.
- The smoking cessation study of Liu, Daniels and Marcus (JASA '09) :

  Q = smoking cessation (0=No, 1=Yes)
  W = weight change
  X = time spent exercising
- Does exercise weaken the association between smoking status and weight gain?

Why CC? Building General Multivariate Distributions

- Joint models for multivariate data.
- If the joint distribution of U_1, U_2, U_3 ($U_i \sim$ Uniform$(0,1)$, $1 \leq i \leq 3$) is modelled using the pair copula model then

$$c(u_1,u_2,u_3) = c_{12}(u_1,u_2)c_{23}(u_2,u_3)c_{13|2}(u_{1|2},u_{3|2};u_2)$$

  where $u_{k|2} = Pr(U_k \leq u_k | U_2 = u_2)$.
- As dimension increases, the bivariate conditional copulas depend on increasing number of variables.

Why CC? Regression-based prediction

- In a bivariate CC model the joint density is

$$h_x(y_1, y_2) = f_x(y_1) g_x(y_2) c_{\theta(x)}(F_x(y_1), G_x(y_2)).$$

- The conditional density of $Y_1 | Y_2 = y_2, X = x$ is

$$h_x(y_1|y_2) = f_x(y_1) c_{\theta(x)}(F_x(y_1), G_x(y_2)).$$

- This can be useful when for each item a subset of the response variables is much easier to measure than the rest.

Propagation of Errors

- Errors can appear from multiple sources:

$$
\begin{aligned}
c_{\theta(x)+\delta_3(x)}(F_x(y_1)+\delta_1(x), G_x(y_2)+\delta_2(x)) &= \overbrace{c_{\theta(x)}(F_x(y_1), G_x(y_2))}^{\text{correct part}} \\
&+ \overbrace{c^{(1,0,0)}_{\theta(x)}(F_x(y_1), G_x(y_2))\ \delta_1(x)}^{\text{error due to } F_x} + \overbrace{c^{(0,1,0)}_{\theta(x)}(F_x(y_1), G_x(y_2))\delta_2(x)}^{\text{error due to } G_x} + \\
&+ \overbrace{c^{(0,0,1)}_{\theta(x)}(F_x(y_1), G_x(y_2))\delta_3(x)}^{\text{error due to } \theta(x)} + \mathcal{O}(||\delta(x)||^2)
\end{aligned}
$$

Estimation of $\theta(x)$ - frequentist approaches

- Acar, Craiu & Yao (Biomcs, 2011) - semiparametric estimation. Parametric marginals, $\theta(x)$ is approximated nonparametrically via local polynomial estimation.
- Veraverbeke, Omelka & Gijbels (SJS, 2011) - nonparametric estimation of the copula and marginals.
- "We observed that the copula estimator may be severely biased if any of the conditional marginal distributions change with the value of the covariate $X = x$" (V., O. & G, 2011)
- Nonparametric estimates in large-ish dimensions d suffer from curse of dimensionality, unless the volume of data is huge.

CC: The SA Condition

- When $d = \dim(X) >> 1$ the curse of dimensionality can be alleviated with dimension reduction models.
- The SA leads to "dimension crushing":

$$C_x(F_x(y_1), G_x(y_2)) = C(F_x(y_1), G_x(y_2))$$

but when is it justifiable?

- Acar, Genest & Nešlehová (JMVA, 2012) - discuss the bias incurred when SA is not justified.
- Acar, Craiu & Yao (EJS, 2013) - Generalized LRT to test a constant or linear null calibration against a general alternative.
- Gijbels, Omelka & Veraverbeke (Statistics, 2016) - nonparametric testing procedures.
- Derumigny & Fermanian (arXiv, 2016) - review of state-of-art and a "work program around SA for the next years".

CC: $d = \dim(X) > 1$

- Sabeti, Wei & Craiu (Stat, 2014) - Bayesian additive CC models.
- Chavez-Demoulin & Vatter (JMVA, 2015) - Generalized additive models.
- Lobato, Lloyd & Lobato (NIPS, 2013) - Gaussian Process models for CC in financial time series.
- Levi & Craiu - Gaussian Process with Single Index Models.

Why a Bayesian approach?

- Joint modelling can be a bit easier as selection of most tuning parameters is automatic and data driven.
- The posterior distribution accounts for all sources of variation (included in the model).
- The Monte Carlo samples from the posterior are used to compute finite sample variance estimates, pointwise credible regions and model selection criteria.
- Priors can be used to favour model sparsity.

Additive Model for Calibration

- $X \in \mathbf{R}^d$ Y_1, Y_2 are continuous r.v.'s, $Y_i \sim \mathcal{N}(\mu_i(X), \sigma_i^2)$, for $i = 1, 2$.
- Jointly,

$$
\begin{aligned}
h_X(Y_1, Y_2) &= \prod_{i=1}^{2} \frac{1}{\sigma_i} \phi\left(\frac{Y_i - \mu_i(X)}{\sigma_i}\right) \\
&\times \quad c\left[\Phi\left(\frac{Y_1 - \mu_1(X)}{\sigma_1}\right), \Phi\left(\frac{Y_2 - \mu_2(X)}{\sigma_2}\right) \middle| \theta(X)\right],
\end{aligned}
$$

where $c(u, v|\theta)$ is the pdf of the conditional copula.

Additive Model for Calibration

- Usually there is little/no information about the shape of $\theta(X)$.
- Generally $\theta(X)$ has a restricted range, so we estimate the calibration function $\eta : \mathbf{R}^d \to \mathbf{R}$ where $g(\theta(x_i)) = \eta(x_i)$ (g is user-specified).
- We assume that

$$\eta(x_1, \ldots, x_d) = \alpha_0 + \sum_{i=1}^{d} \eta_i^*(x_i).$$

- Additivity is not preserved when changing dependence measure.

Additive Model for Calibration

- Sabeti, Wang & Craiu (Stat, 2013) use an AM:
  - Each η_i^* has the form:

$$\eta_i^*(x_i) = \sum_{j=1}^{3} \alpha_j^{(i)} x_i^j + \sum_{k=1}^{K^{(i)}} \psi_k^{(i)} (x_i - \gamma_k^{(i)})_+^3.$$

  - Number and location of knots $\{\gamma_1, \ldots, \gamma_{K^{(i)}}\}$ is important.
  - The knot-related choices are data driven.
- Partition the range of X_i into K_{max} intervals, $I_k^{(i)}$, and introduce auxiliary variables

$$\zeta_k^{(i)} = \begin{cases} 1 & \text{if there is a knot } \gamma_k^{(i)} \text{ in } I_k^{(i)} \text{ and } \psi_k^{(i)} \neq 0 \\ 0 & \text{if there is no knot in } I_k^{(i)} \text{ and } \psi_k = 0 \end{cases}$$

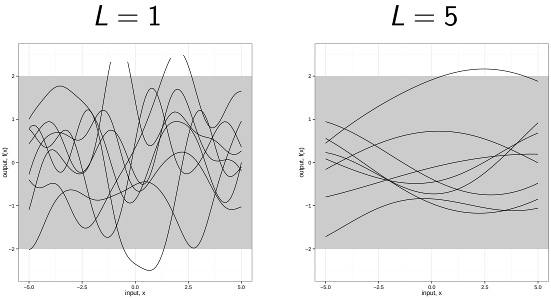
- $\eta_i^*(x_i) = \sum_{j=1}^{3} \alpha_j^{(i)} x_i^j + \sum_{k=1}^{K_{max}} \psi_k^{(i)} \zeta_k^{(i)} (x_i - \gamma_k^{(i)})^3{}_+,$

Gaussian Process Prior for CC

- GP is a flexible method when errors are reasonably approximated by Gaussians.
- GP for marginals when means could be nonlinear functions of X.
- GP for calibration function could be used in conjunction with other marginal models.
- Vanilla GP is not helping with the curse of dimensionality and can be expensive when n is large so modifications are needed.

Gaussian Process Prior

- GP prior for smooth f without specifying the form of f.
- For $x \in [-5, 5]^n$, consider $f \sim N_n(0, K(x, x))$ where $K_{ij}(x, x) = k(x_i, x_j)$ and $f_i = f(x_i)$



- Random functions f generated from a GP prior when $n = 100$
- $\mathrm{Cov}(f(x_i), f(x_j)) = k(x_i, x_j) = \exp\{-0.5 * \frac{|x_i - x_j|}{L}^2\}$.

Gaussian Process Estimation

- Observe $\{y_i : 1 \leq i \leq n\}$ noisy realizations of $f(x_i)_{i=1,n}$, $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$.
- When interested in predicting $f^* = (f(x_j^*))_{j=1,q}$ use

$$\left(\begin{array}{c} y \\ f^* \end{array} \right) \sim N_{n+q}\left(\mathbf{0}, \left[\begin{array}{cc} K(x,x) + \sigma^2 \mathbf{I}_n & K(x,x^*) \\ K(x,x^*) & K(x^*,x^*) \end{array} \right]\right)$$

- The conditional distribution of f^* is Gaussian with

$$\mathrm{E}(f^*|y) = K(x^*,x) \overbrace{[K(x,x) + \sigma^2 \mathbf{I}_q]^{-1}}^{\text{expensive for large } n} y$$

$$V(f^*|y) = K(x^*,x^*) - K(x^*,x) \underbrace{[K(x,x) + \sigma^2 \mathbf{I}_q]^{-1}}_{\text{expensive for large } n} K(x,x^*)$$

Computational challenges for GP

- When n is large the computation effort is prohibitive so we adopt a sparse GP approach.
- The information about f in the data is funnelled using a smaller sample of size $m << n$ of inducing (or latent) variables $\tilde{x}_g$, $1 \le g \le m$.
- Let $\tilde{\mathbf{f}} = (f(\tilde{x}_1), \ldots, f(\tilde{x}_m))^T$ we assume $\tilde{\mathbf{f}} \sim N(0, K(\tilde{X}, \tilde{X})$.
- The (Gaussian) conditional distribution of $\tilde{f}|X, \tilde{X}$ involves $K(X, \tilde{X})$ and the inverse of a $m \times m$ matrix.
- Integrating out $\tilde{f}$ yields the (Gaussian) conditional distribution of $f^*|X^*, X$ with similar computational burden.

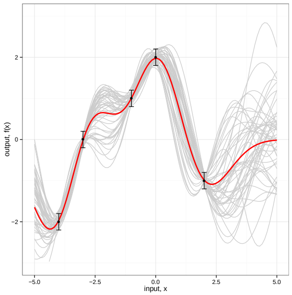
Gaussian Process Estimation

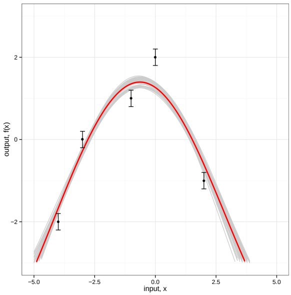
- Suppose we know that $f : \mathbf{R} \to \mathbf{R}$ and

$$(-4, -3, -1, 0, 2) \xrightarrow{f} (-2, 0, 1, 2, -1)$$

- $y_i | f_i \sim N(f_i, \sigma^2), \quad f_i = f(x_i),\ 1 \leq i \leq 5$

$L = 1$ $\qquad$ $L = 5$





Samples from the posterior distribution of $f^*|y$, when $\sigma^2 = 0.1$.

Modelling η when $d > 1$

- Assume that $\theta(X_i) = g^{-1}(f(X_i))$ and

$$\mathbf{f} = (f(X_1), f(X_2), \ldots, f(X_n))^T \sim \mathcal{N}(0, K(X, X; \mathbf{w})),$$

- The (i, j)th element of matrix $K(X, X; \mathbf{w})$ is

$$k(x_i, x_j; \mathbf{w}) = e^{w_0} \exp\left[-\sum_{s=1}^{d} \frac{(x_{is} - x_{js})^2}{e^{w_s}}\right].$$

- The covariance between outputs is defined as a function of inputs.
- The parameters $\mathbf{w}$ in the covariance function k determine distance between inputs with a significant difference in the outputs.

Modelling η when $d > 1$

- When $X_i \in \mathbf{R}^d$ a full GP model for the CC has $1 + d$ parameters.
- We consider instead the SIM model
$$f(X) = f(\beta^T X).$$
- GP-SIM model is invariant to nonlinear one-to-one transformations $\tau(\theta)$.
- The parameter β is unidentifiable up to a constant so we assume $||\beta|| = 1$.
- Marginals are fitted also using GP-SIM models.

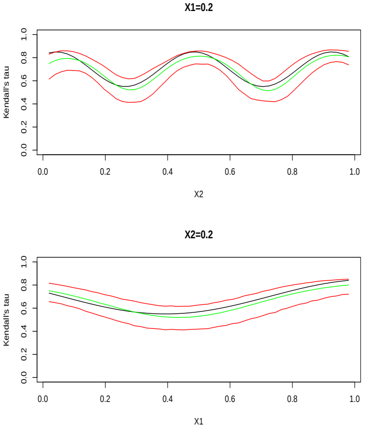
Proof of concept

Sc1 $f_1(x) = 0.6\sin(5x_1) - 0.9\sin(2x_2),$
$f_2(x) = 0.6\sin(3x_1 + 5x_2),$
$\tau(x) = 0.7 + 0.15\sin(15x^T\beta)$
$\beta = (1,3)^T/\sqrt{10},\ \sigma_1 = \sigma_2 = 0.2\ \ n = 400$

	Clayton			Frank			Gaussian			Clayton SA		
Scenario	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$	$\sqrt{\text{IBias}^2}$	$\sqrt{\text{IVar}}$	$\sqrt{\text{IMSE}}$
Sc1	**0.0223**	0.0556	**0.0599**	0.0491	0.0714	0.0867	0.0664	0.0741	0.0995	0.1071	**0.0133**	0.1079

Integrated error for the estimator of $\tau(x)$.

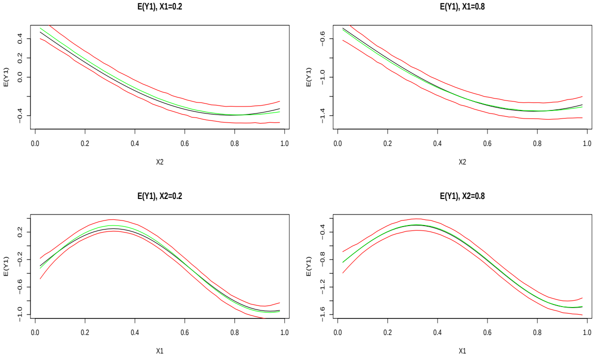
Estimation of $\tau(x)$



Prediction performance

- If $y_i|x \sim N(\mu_i(x), \sigma_i^2)$, $i = 1, 2$ then

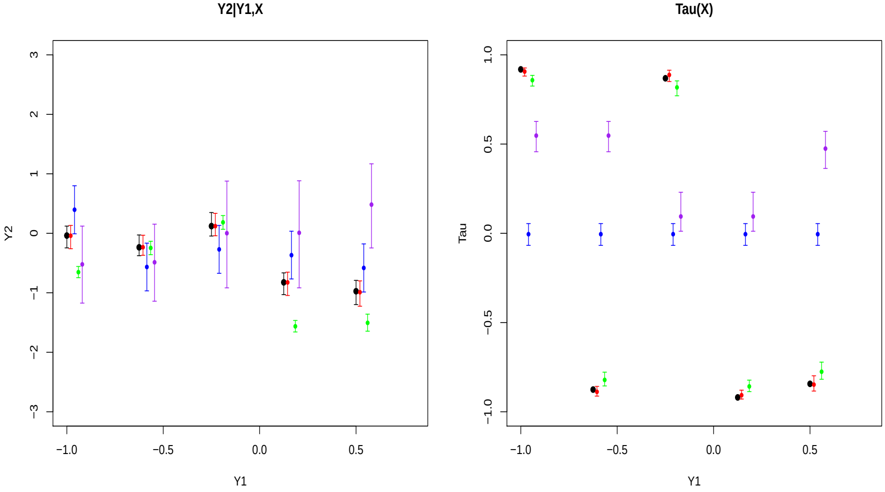
$$E_x[Y_1|Y_2 = y_2] = \mu_1(x) + \sigma_1 \int_0^1 \Phi^{-1}(z) c_{\theta(x)} \left(z, \Phi\left(\frac{y_2 - \mu_2(x)}{\sigma_2} \right) \right) dz.$$



Model misspecification effects

- Marginals:
  - $f_1(x) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
  - $f_2(x) = 0.6\sin(3x_1 + 5x_2)$
  - $\sigma_1 = \sigma_2 = 0.2$
- Copula:
  - $\beta = (1,3)/\sqrt{10}$
  - $\tau(x) = 0.7 + 0.15\sin(5x^T\beta)$
  - Frank copula
- Y1 and X test points:
  - $Y_1 = -1.5,\ X = (0.3, 0.3)$
  - $Y_1 = -1.0,\ X = (0.3, 0.7)$
  - $Y_1 = -0.5,\ X = (0.7, 0.3)$
  - $Y_1 = 0.0,\ X = (0.7, 0.7)$
  - $Y_1 = 0.5,\ X = (0.5, 0.5)$

Model misspecification effects



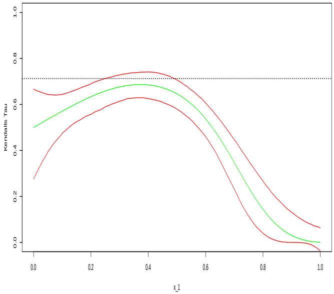
Black - true; Red - Correct model; Blue - Correct copula with SA
Green - Wrong copula; Purple - Correct copula with missing covariate

Model misspecification effects

- Marginals
  - $f_1(x) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
  - $f_2(x) = 0.6\sin(3x_1 + 5x_2)$
  - $\sigma_1 = \sigma_2 = 0.2$, $X_1 \perp X_2$.
- Copula: $\tau(x) = 0.71$
- Model:
  - A nonparametric model for marginals and CC based on only x_1.

Model misspecification effects

- Marginals
  - $f_1(x) = 0.6\sin(5x_1) - 0.9\sin(2x_2)$
  - $f_2(x) = 0.6\sin(3x_1 + 5x_2)$
  - $\sigma_1 = \sigma_2 = 0.2$, $X_1 \perp X_2$.
- Copula: $\tau(x) = 0.71$
- Model:
  - A nonparametric model for marginals and CC based on only x_1.



Model misspecification effects

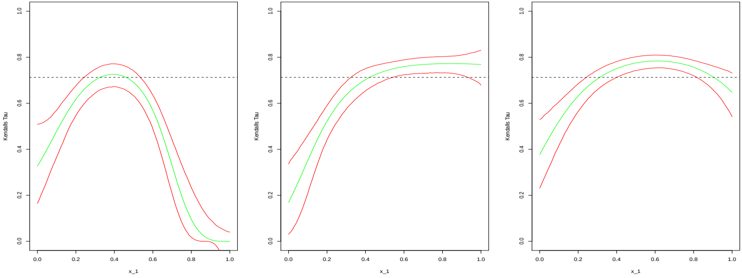
- Suppose $E[Y_i|X_1, X_2] = f_i(\alpha_i X_1 + \beta_i X_2)$, i=1,2.
- Then
$$f_i(x_1, x_2) = \overbrace{f_i(x_1, 0)}^{\text{fit}} + f_i^{(0,1)}(x_1, 0)\beta_i x_2 + f_i^{(0,2)}(x_1, 0)\frac{\beta_i^2 x_2^2}{2} + \mathcal{O}(||x_2^3||),$$
$i = 1, 2$
- The marginal residuals still contain high-order information about x_1 which varies with the distribution of x_2.

Model misspecification effects

- Suppose $E[Y_i|X_1, X_2] = f_i(\alpha_i X_1 + \beta_i X_2)$, i=1,2.
- Then

$$f_i(x_1, x_2) = \overbrace{f_i(x_1, 0)}^{\text{fit}} + f_i^{(0,1)}(x_1, 0)\beta_i x_2 + f_i^{(0,2)}(x_1, 0)\frac{\beta_i^2 x_2^2}{2} + \mathcal{O}(||x_2^3||),$$
$i = 1, 2$

- The marginal residuals still contain high-order information about x_1 which varies with the distribution of x_2.



Three Model Selection Problems

- Choice of copula family.
- Choice of calibration
  - Simplifying Assumption or not?
  - AM or GP-SIM?
- Covariate selection.

CV Marginal Likelihood (CVML)

- Calculates the average (over parameter values) prediction potential for model $\mathcal{M}$ via

$$\text{CVML}(\mathcal{M}) = \sum_{i=1}^{n} \log\left(P(Y_{1i}, Y_{2i} | \mathcal{D}_{-i}, \mathcal{M})\right),$$

- $\mathcal{D}_{-i}$ is the data set from which the ith observation has been removed.
- Estimate CVML using

$$E_{\pi}\left[P(Y_{1i}, Y_{2i} | \omega, \mathcal{M})^{-1}\right] = P(Y_{1i}, Y_{2i} | \mathcal{D}_{-i}, \mathcal{M})^{-1}$$

where ω represents the vector of all the parameters and latent variables in the model.

Watanabe-Akaike Information Criterion

- An alternative approach to estimating the log pointwise predictive density (Watanabe, JMLR '10; Vehtari, Gelman & Gabry, Stat. Comput, '16).
- WAIC is defined as

$$\text{WAIC}(\mathcal{M}) = -2\text{fit}(\mathcal{M}) + 2\text{p}(\mathcal{M})$$

  - $\text{fit}(\mathcal{M}) = \sum_{i=1}^{N} \log \mathrm{E}_{\pi}\left[P(y_{1i}, y_{2i}|\omega, \mathcal{M})\right]$
  - $\text{p}(\mathcal{M}) = \sum_{i=1}^{N} \mathrm{Var}_{\pi}[\log P(y_{1i}, y_{2i}|\omega, \mathcal{M})]$.
- $\text{fit}(\mathcal{M})$ and the penalty $\text{p}(\mathcal{M})$ are computed using the posterior samples.

Another Two Scenarios

Sc3

$$\left.\begin{array}{c} f_1(x) = 0.6\sin(5x_1) - 0.9\sin(2x_2) \\ f_2(x) = 0.6\sin(3x_1 + 5x_2) \\ \tau(x) = 0.5 \\ \sigma_1 = \sigma_2 = 0.2 \end{array}\right\}$$

SA is true

Sc4

$$\left.\begin{array}{c} f_1(x) = 0.6\sin(5x_1) - 0.9\sin(2x_2) \\ f_2(x) = 0.6\sin(3x_1 + 5x_2) \\ \eta(x) = 1 + 0.7\sin(3x_1^3) - 0.5\cos(6x_2^2) \\ \sigma_1 = \sigma_2 = 0.2 \end{array}\right\}$$

AM model is true

Copula Selection - Results

	Frank		Gaussian		Clayton SA	
Scenario	CVML	WAIC	CVML	WAIC	CVML	WAIC
Sc1	100%	100%	100%	100%	98%	98%
Sc4	100%	100%	100%	100%	100%	100%

Calibration Selection - Results

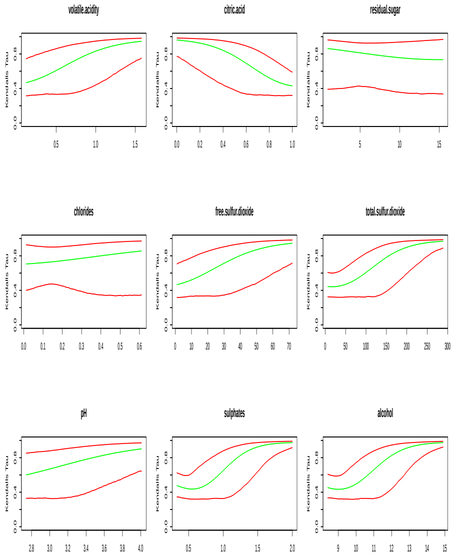
Scenario	Clayton		Frank		Gaussian	
	CVML	WAIC	CVML	WAIC	CVML	WAIC
Sc3	78%	78%	100%	100%	100%	100%

Scenario	CVML	WAIC
Sc 1	96%	96%
Sc4	94%	94%

Red Wine

- 11 Physiochemical properties of 1599 varieties of red "Vinho Verde" Portuguese wine.
- We consider the dependence between fixed acidity and density. The former is strongly associated with quality of wine, while the latter is used as a measure of grape quality.
- Covariates: volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, alcohol.

Red Wine


volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
pH
sulphates
alcohol
Kendalls Tau

Red Wine

	2.5% quantile	mean	97.5% quantile
volatile acidity	0.142	0.274	0.406
citric acid	-0.406	-0.320	-0.159
residual sugar	-0.315	-0.089	0.244
chlorides	-0.303	0.024	0.264
free sulfur dioxide	0.123	0.308	0.514
total sulfur dioxide	0.219	0.396	0.584
pH	0.052	0.166	0.324
sulphates	0.335	0.478	0.611
alcohol	0.381	0.467	0.549

Red wine data: posterior estimates and credible regions for β.

Conclusions

- Flexible modelling for CC is necessary.
- CC models can be useful in prediction when the dependence is strong.
- The SA is difficult to detect without a properly calibrated penalty for model complexity.
- Impact of model misspecification can be significant. Robust CC's?
- How to extend these methods to more than two responses (e.g. vines).
- CC-based models in clustering, adaptive MCMC, spatial models, statistical genetics.