Risk Analysis DOI: 10.1111/risa.13695

Vine Regression with Bayes Nets: A Critical Comparison with Traditional Approaches Based on a Case Study on the Effects of Breastfeeding on IQ

Roger M. Cooke,[1,2,*] Harry Joe,[3] and Bo Chang[3]

Regular vines (R-vines) copulas build high dimensional joint densities from arbitrary one-dimensional margins and (conditional) bivariate copula densities. Vine densities enable the computation of all conditional distributions, though the calculations can be numerically intensive. Saturated continuous nonparametric Bayes nets (CNPBN) are regular vines. Computing regression functions from the vine copula density is termed vine regression. The epicycles of regression–including/excluding covariates, interactions, higher order terms, multicollinearity, model fit, transformations, heteroscedasticity, bias–are dispelled. One simply computes the regressions from the vine copula density. Only the question of finding an adequate vine copula remains. Vine regression is applied to a data set from the National Longitudinal Study of Youth relating breastfeeding to *IQ*. The expected effects of breastfeeding on *IQ* depend on *IQ*, on the baseline level of breastfeeding, on the duration of additional breastfeeding and on the values of other covariates. A child given two weeks breastfeeding can expect to increase his/her *IQ* by 1.5–2 *IQ* points by adding 10 weeks of breastfeeding, depending on values of other covariates. A child given two years breastfeeding can expect to gain from 0.48–0.65 *IQ* points from 10 additional weeks. Adding 10 weeks breastfeeding to each of the 3,179 children in this data set has a net present value $50,700,000 according to the Bayes net, compared to $29,000,000 according to the linear regression.

KEY WORDS: Bayes net; breastfeeding; copula; Gaussian copula; heteroscedasticity; IQ; multivariate regression; National Longitudinal Study of Youth; regression heuristics; Regular vine; vine copula

1. INTRODUCTION

A Regular vine (R-vine) copula (Bedford & Cooke, 2002; Cooke 1997) is a tool for constructing high dimensional distributions with dependence. One-dimensional margins can be fitted from data, the dependence structure is represented by sets of bivariate and conditional bivariate copulas. A Wikipedia page provides an informal introduction (https://en.wikipedia.org/wiki/Vine_copula). For definitions and properties of vines, see Joe (2014), Kurowicka and Cooke (2006), Kurowicka and Joe (2010), for their historical origins see Cooke, Joe, and Aas (2010) and Joe (1994). Vines are most actively employed in financial mathematics (Aas & Berg, 2009; Aas, Czado, Frigessi, & Bakken, 2009; Chollete, Heinen, & Valdesogo, 2009; Czado, Brechmann, & Gruber, 2013; Fischer, Köck, Schlüter, & Weigert, 2009; Jaworski, Durante, & Härdle, 2013; Low, Alcock, Faff, & Brailsford, 2013). Software has been developed at the TU Munich (Brechmann & Schepsmeier, 2013; Nagler, Schepsmeier,

[1] Resources for the Future, Washington, DC, USA.

[2] Department of Mathematics, Delft University of Technology, Delft, The Netherlands.

[3] Deptartment of StatisticsUniversity of British Columbia, Vancouver, Canada.

[*] Address correspondence to Roger M. Cooke, Resources for the Future, Washington, DC, USA; cooke@rff.org

0272-4332/21/0100-0001$22.00/1

Stoeber, Brechmann, & Graeler, 2019), TU Delft (Hanea, Kurowicka, Cooke, & Ababei, 2010), and the University of British Columbia (Joe, 2014). For the special case of (Gaussian) Bayes nets, software is available, and it is capable of handling very large problems (https://www.tudelft.nl/ewi/over/de/faculteit/afdelingen/applied/mathematics/applied/probability/risk/software/uninet/).

The number of labeled R-vines on *n* variables is quite large (Cooke, Kurowicka, & Wilson, 2015; Morales, 2009):

$$\binom{n}{2}(n-2)!2^{\binom{n-2}{2}}, \tag{1}$$

and any absolutely continuous distribution on *n* variables may be represented on any Regular vine with density $f_{1,2\ldots n}$ written as a product of one-dimensional marginal densities $f_1 \ldots f_n$ and copula densities (Bedford & Cooke, 2001):

$$f_{1,2,\ldots,n}(x_1,\ldots,x_n) = f_1(x_1)\ldots f_n(x_n)\prod_{e\in\upsilon} c_{e_1,e_2;}D_{(e)}.(2)$$

Here, edge $e = \{e1,e2,D(e)\}$ in edge set ν has conditioning set *D(e)* and conditioned variables e_1, e_2. The copula density function $c_{e1,e2\ ;D(e)}$ may depend on the conditioning set *D(e)*, that is, a different copula function may be used for different values of *D(e)*. The "simplifying assumption" that the copula density functions do not depend on the values of the conditioning variables is often invoked, resulting in "simplified R-vines" (Acar, Genest, & Neshlehova, 2012; Hobaek Haff, Aas, & Frigessi, 2010; Stoeber, Joe, & Czado, 2013). For simplified vines the bivariate copulas in trees 2 and higher depend on the values of the conditioning variables only through the conditional cumulative distribution functions of (e_1,e_2). It is not the case that any absolutely continuous distribution can be represented in the above form on any simplified R-vine; some simplified R-vines will fit a multivariate data set better than others. A Gaussian R-vine, where the copulas on the edges of the vine are Gaussian satisfies the simplifying assumption and can represent any absolutely continuous distribution with a multivariate Gaussian copula.

A simplified R-vine requires estimation of *n(n-1)/2* bivariate copulas and allows empirical univariate margins. Compared to other parametric families of multivariate distribution functions, those represented on R-vines confer enormous modeling flexibility with a manageable number of estimated parameters. The conditional distribution of any set of variables given any disjoint set of variables can be computed, hence also the conditional expectations. In other words, ALL regression functions can be computed. Vine copula density estimation thus affords a viable alternative to stipulating an algebraic form for regression functions and estimating their coefficients from data. Choosing an algebraic form raises questions like:

- Can or should some potential predictors or covariates be excluded from predicting a response variable of interest?
- Should some covariates be transformed?
- Is multicollinearity a concern?
- Should higher order terms be included?
- Should interactions be included?
- Is the error variance constant (homoscedastic)?
- If homoscedasticity fails, how should we estimate the error variance?
- Without known functional form for the error variance how should we estimate regression coefficients?
- How should we evaluate model fit?

Such concerns are denoted here the epicycles of regression (see also, Sala-I-Martin, 1997). More detail on vine regression is found in Chang and Joe (2019), Kraus and Czado (2017), and Parsa and Klugman (2011).

This article explores two uses of vine regression when one of the observed variables is to be predicted. First (1) using vine models to fit or to smooth data and (2) using the vine copula density to compute regression functions. For (1), the distinction between fitting and smoothing is not sharp; fitting usually minimizes some measure of lack of fit, whereas smoothing tries to reveal underlying structure by blurring out detail. The vine copula density may be used to compute regression functions. For (2) because of Equation (2), we can draw arbitrarily many samples from a wide variety of multivariate distributions for which the exact regression functions are known. This confers the possibility of ground truthing the heuristics used to address the above epicycles. That is, we compare different regression heuristics with the true regression functions.

This article illustrates both possibilities and focuses on Gaussian vines represented as Bayes nets. These are not intended to fit the data, and they may miss features like tail dependence and asymmetry (Joe, Li, & Nikoloulopoulos, 2010). On the other hand, they often do a reasonable job of representing rank correlations and enabling an intuitive graphical representation with fast analytical

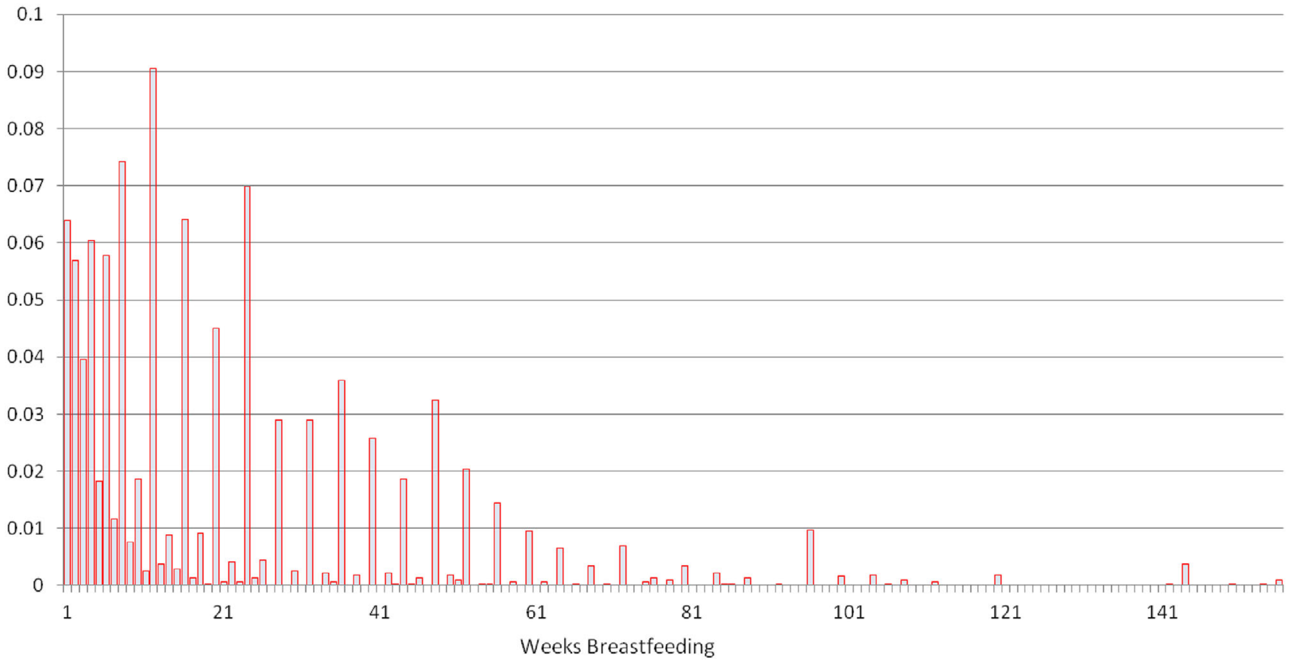

Fig 1. Histogram of number of weeks breastfeeding in NLSY data.

conditionalization. The accuracy of a Bayes net can be checked by comparing outcomes with those from good fitting R-vine. Additional detail on vine regression applied to the effect of breastfeeding on *IQ* is found in Cooke, Joe and Chang (2015, 2020).

Section 2 introduces a data set from the National Longitudinal Study of Youth (NLSY) for studying the relation between breastfeeding and IQ, in the presence of other covariates. Section 3 presents a Gaussian smoothed emulation of the data, Section 4 compares various regression heuristics with the "ground truth" obtained by conditionalization. Section 5 discusses the question of a good fitting R-vine for this data set, and Section 6 concludes.

2. NLSY DATA

There is a great deal of policy interest on the effects of breastfeeding both in the developed and the developing world and much controversy surrounds attempts to isolate this effect (for a review see Horta & Victora, 2013). The NLSY is perhaps the most often cited data set in this discussion.

To study the effect of breastfeeding duration, we down select to children who were ever breast fed and focus on (roughly) continuous covariates with mild and independent censoring. Retaining only data without censoring, a data set of 3,179 samples is obtained for a child's IQ, measured by the Peabody pictorial visual test, a nonverbal test for IQ, usually taken at age 8–10. The explanatory variables (covariates) are weeks breastfeeding (BFW), Mother's IQ measured by the armed forces qualification test ($MAFQT$, not scaled as an IQ test, but closely correlated with IQ, usually taken at age 18), family income at child's birth (INC), Mother's highest completed grade of schooling ($MGRADE$), Mothers age at child's birth ($MAGE$), and child's year of birth ($CBIRTH$). The goal is to quantify the effect of BFW on IQ in the presence of these covariates.

The reported number of weeks of breastfeeding (Fig. 1) range from 1 to 156. 2678 of the 3,179 reported weeks breastfeeding are even, presumably a spurious mnemonic artifact. Among the odd numbers, only 59% are above one week. Many of the one-week entries may indicate a failed attempt at breastfeeding, thereby conflating the effect of breastfeeding duration with the effect of ever versus never breastfed. On the other hand, the effect of additional breastfeeding is strongest for children with the smallest duration of breastfeeding (see Fig. 4). Hence, restricting to children with at least two weeks breastfeeding probably leads to an under estimate of the effect of duration of breastfeeding, whereas including children with less than two weeks breastfeeding probably leads to an over estimate. In computing the

Table I. Multiple linear regression for NLSY data, see text for explanation of covariates. R^2 from the linear regression is 0.236

	Coefficients	*SE*	*t* Stat	*p*-value	Lower 95%	Upper 95%
CBIRTH	−0.065	0.124	−0.522	0.602	−0.308	0.179
BFW	0.051	0.013	4.093	0.000	0.027	0.076
MAGE	−0.597	0.453	−1.320	0.187	−1.485	0.290
MGRADE	0.589	0.128	4.614	0.000	0.339	0.839
MAFQT	0.262	0.012	22.513	0.000	0.239	0.285
LnINC	17.986	11.414	1.576	0.115	−4.395	40.366

Table II. Rank correlation matrix for NLSY data

	MAFQT	*INC*	*MGRADE*	*CBIRTH*	*MAGE*	*BFW*	*IQ*
MAFQT	1.00	0.46	0.52	0.21	0.22	0.24	0.50
INC	0.46	1.00	0.46	0.62	0.62	0.17	0.32
MGRADE	0.52	0.46	1.00	0.26	0.28	0.18	0.31
CBIRTH	0.21	0.62	0.26	1.00	0.93	0.17	0.14
MAGE	0.22	0.62	0.28	0.93	1.00	0.19	0.15
BFW	0.24	0.17	0.18	0.17	0.19	1.00	0.18
IQ	0.50	0.32	0.31	0.14	0.15	0.18	1.00

effect of breastfeeding duration on *IQ* (Table V) both options are given.

The output of a multiple linear regression is given in Table I (the fat tailed covariate INC is logged).

The rank correlation matrix of the NLSY data displays some substantial dependence among the independent variables, which reduces the precision of the coefficients' estimates.

3. VINE REGRESSION

A continuous nonparametric Bayes net (CNPBN) is constructed using the one-dimensional empirical distributions and dependence based on Table II. Since the (conditional) copulas in CNPBNs are all Gaussian, they are parametrized by (conditional) rank correlation. *IQ* is the unique sink node; all arcs are incoming meaning that all influences are toward *IQ* (see Fig. 2). *MAFQT* is the first root, all its arcs are outgoing, and all rank correlations associated with these arcs are unconditional. The influences are interpreted as going from *MAFQT* to the other nodes. The second root is *INC*, all its arcs except that from *MFQT* have *INC* as their source, and these arcs are associated with conditional rank correlations given *MAFQT*. The third root is *MGRADE*, and its source arcs are associated with conditional rank correlations given *MAFQT* and *INC*. Proceeding in this way, *CBIRTH* is the source of three arcs, *MAGE* is the source of two arcs, *BFW* is the source of one arc; its correlation reflects the influence of *BFW* on *IQ* after the influence of other nodes is removed. The theory of regular R-vines tells us that these (conditional) copulas together with the one-dimensional margins uniquely determines the joint density and that the (conditional) copulas are algebraically independent. For details on CNPBNs and their relation to R-vines, see Kurowicka and Cooke (2006), and for R-vines see Bedford and Cooke (2002) and Cooke (1997).

If p_j denotes a partial rank correlation of variable j with sink node *IQ* in the conditioned set, $[1 - \Pi_j (1 - p_j^2)]^{0.5}$ is the multiple rank correlation of *IQ* on all the other variables (Kurowicka & Cooke, 2006). Its square is the *vine copula-*R^2. Reading the values of partial rank correlations from Fig. 2 yields *vine copula-*$R^2 = 0.2416$, which is nearly equal to the R^2 from linear regression (0.236). In other words, the fraction of variance of *IQ* explained by other variables in the original predictor space is nearly equal to the fraction of explained variance after transforming all variables to uniform with the probability integral transformation (the copula space).

The partial rank correlations are derived from the multivariate normal distributions whose rank correlation matrix is closest to the empirical rank correlation matrix of the data in Table II. More precisely, we transform each variable X_i with *CDF* F_i to

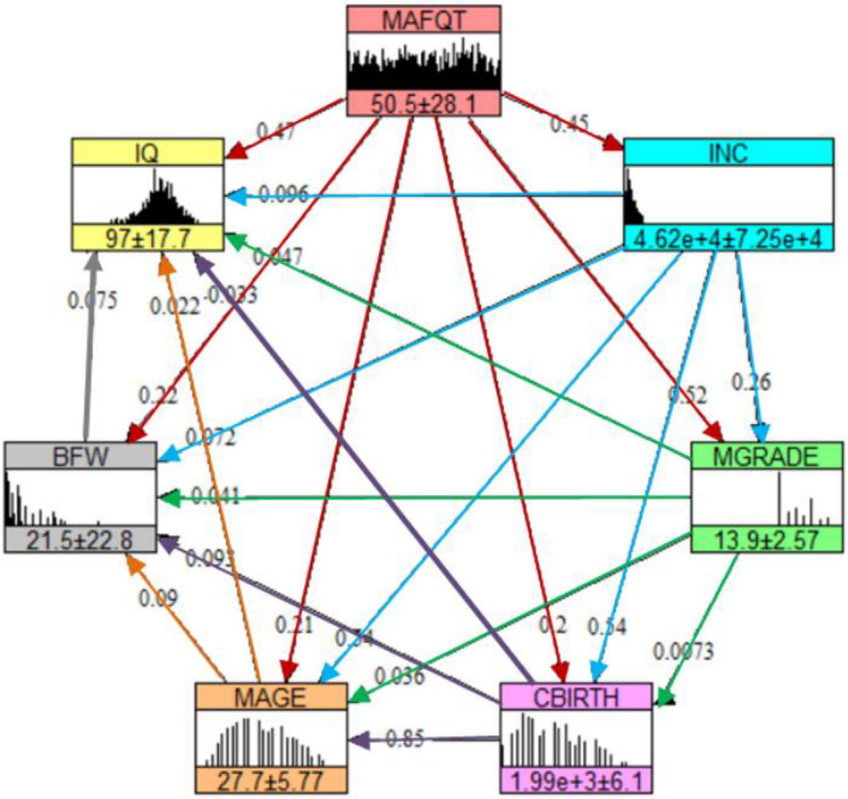

Fig 2. Bayes net for NLSY data. Each node shows the empirical distribution of the corresponding variable as well as its mean and standard deviation (±). The Bayes net is built of nested trees, each tree with one root, that is a node with only outgoing arcs in that tree. The first tree is red, the second is blue, third is green, fourth is purple, fifth is orange, and sixth is black. Arcs are associated with (conditional) partial rank correlations. The red arcs from the first tree are unconditional, the blue arcs from the second tree are conditional on MAFQT, the green arcs from the third tree are conditional on MAFQT and INC, the purple arcs from the fourth tree are conditional on MAFQT, INC, and MGRADE, the orange arcs from the fifth tree are conditional on MAFQT, INC, MGRADE, and CBIRTH, the black arc from sixth tree is conditional on all preceding nodes. These partial rank correlations together with the one-dimensional margins and the copula uniquely determine the joint distribution.

Table III. Rank Correlation Matrix from Gaussian Bayes net left of "\" and rank correlations from NLSY data (right of "\")

	MAFQT	*INC*	*MGRADE*	*CBIRTH*	*MAGE*	*BFW*	*IQ*
MAFQT	1	0.45\0.46	0.52\0.52	0.2\0.21	0.21\0.22	0.22\0.24	0.47\0.5
INC		1	0.44\0.46	0.57\0.62	0.57\0.62	0.16\0.17	0.29\0.32
MGRADE			1	0.23\0.26	0.26\0.28	0.16\0.18	0.31\0.31
CBIRTH				1	0.9\0.93	0.16\0.17	0.12\0.14
MAGE					1	0.19\0.19	0.14\0.15
BFW						1	0.18\0.18
IQ							*1*

standard normal as $Z_i = \Phi^{-1} F_i(X_i)$, where Φ is the standard normal CDF. $\mathbf{Z} = (Z_1,\ldots Z_n)$ is not multivariate normal, but we consider a multivariate normal vector ***Z'*** with the same covariance matrix as ***Z***. Fig. 2 shows partial rank correlations of ***Z'***; the rank correlation matrix of ***Z'*** is given in Table III. It can be shown that the partials in Fig. 2 uniquely determine the rank correlation matrix (Bedford & Cooke, 2002). Together with the one-dimensional margins and Gaussian copula with associated rank correlations assigned to each arc, these partials uniquely determine the joint distribution (Kurowicka & Cooke, 2006). The joint distribution of $(F_1^{-1}\Phi(Z'_1),\ldots F_n^{-1}\Phi(Z'_n))$ is called the *Gaussian smoothing* of $(X_1,\ldots X_n)$. It replaces the original copula of $X_1,\ldots X_n$ with the closest Gaussian copula.

Using the Gaussian copula, with any given vector of covariate values, we may sample the conditional mean of *IQ* given the covariate values. Doing this (based on 32,000 conditional samples) for each of the 3,179 individuals in the data set, a Gaussian smoothed predictor of *IQ* is found, denoted $E(IQ|\mathbf{X})$. Table IV compares the mean square error and mean absolute deviation of the Gaussian smoothing prediction $(E(IQ|\mathbf{X}))$ and the linear prediction with

Table IV. Mean square error and mean absolute deviation of the Gaussian smoothing prediction (*E(IQ|**X**)*) and the linear prediction with coefficients from Table I, applied to the NLSY data

	*E(IQ|**X**)*	Linear Prediction
RMSE	15.45	15.46
MAD	11.59	11.59

coefficients from Table I, applied to the NLSY data. Over the whole data set the differences are small, but for given values of $X = x$, we shall see that the differences can be large.

3.1. The effect of breastfeeding duration on IQ

It is helpful to reflect on the meaning of "the effect of breastfeeding duration on *IQ*." If we conditionalize on one value *b* of *BFW*, then the expectation of *IQ* given *BFW* = *b*, *E(IQ* | *BFW* = *b)*, will be confounded by all the other covariates which are correlated with *BFW*. This would answer a question like *"given an individual about whom we know only that BFW = b, what do we expect his IQ to be?"* Indeed, *BFW* = *b* also tells us something about the mother's age and the family's income, and so on and this should influence our expectation of *IQ*. With linear regression the increment in expected *IQ* from *b* additional weeks breastfeeding is independent of the values of the other covariates. With vine regression that need not be the case.

One is often interested in a different question: "*If we change only the BFW for an individual, how might that affect that individual's IQ?"* When we change the value of *BFW*, we do not change the family's income or the mother's highest grade of schooling, and so on. Putting $\boldsymbol{X}$ = (*Cbirth, Mage, Mgrade Mafqt, Inc, BFW)*, with possible value $\boldsymbol{x}$, then $E(IQ \mid \boldsymbol{X} = \boldsymbol{x})$ gives the expected *IQ* for an individual with covariate values $\boldsymbol{x}$. The answer to the latter question is found by considering the expected difference in *IQ* for individual $\boldsymbol{x}$ and another individual identical to $\boldsymbol{x}$ except that BFW has been augmented by $\delta > 0$, written $\boldsymbol{x} \setminus x_{BFW} + \delta$. The effect of BFW on *IQ* is then found by integrating this expected difference over values of $\boldsymbol{X}$:

$$\text{Effect BFW on } IQ \;= EXP(1/\delta)[E(IQ|X\backslash X_{BFW} + \delta) - E\,(IQ|X)]\,. \qquad (3)$$

In other words, we integrate the scaled difference of two regression functions which differ only in that one has δ weeks more breastfeeding than the other. Obvious generalizations of (3) would enable joint regression (say of *IQ and INC)* on multivariate effects (say *BFW* and *MGRADE*). These conditionalizations are readily handled with Bayes nets.

Fig. 3 shows the *CDFs* of *IQ, E(IQ* | ***X*** *)*, and also the *CDF* of the linear regression predictor of *IQ* from Table I. *E(IQ* | ***X*****)** and the linear regression predictor are comparable, except on the low *IQ* end.

The linear predictor assumes that the effect of breastfeeding on *IQ* is linear; one additional week breastfeeding adds 0.05 *IQ* points, adding 25 years of breastfeeding adds 65 *IQ* points. By varying δ in (3), vine regression avoids such implausible predictions. Table V shows the effect of breastfeeding duration on *IQ* for values of δ from 1 to 25. To compare with the linear predictor, the effect scaled per week is given. Results of including and excluding individuals with *BFW* = 1 are also shown. The weekly effect approaches the linear predictor value of 0.05 as the number of additional weeks increases.

In vine regression, the effect of adding δ weeks of breastfeeding to a baseline breastfeeding level plateaus as δ increases and as the baseline increases, as is eminently reasonable. In consequence, the effect for low baselines and low δ is higher than the linear model predicts. The combined action of these factors, together with the marginal distribution of *BFW* (Fig. 1) leads the linear model to under predict the benefit to the population of adding, say, 10 weeks breastfeeding to every child in the data base, relative to the Bayes net. To illustrate, we use the estimate of Grosse et al. (2002) that each full scale (reading and math) *IQ* point lost reduces future work productivity by 1.76–2.38%, monetized by Gould (2009) at $17,815 in discounted lifetime earnings.

Extending this calculation to the whole NSLY data set, adding 10 weeks breastfeeding to each of the 3,179 children would add $50,700,000 according to the Bayes net, but only $29,000,000 according to the linear regression.

The differences of the regression functions *E*(*IQ* |***X***) and *E*(*IQ* |***X***\ *BFW* + δ) scattered against *BFW* (for *BFW* > *1*) show that the effect of additional weeks of breastfeeding is greatest for low values of *BFW*. Fig. 4 shows the results for δ = 10 and 20. A child given two weeks breastfeeding can expect to increase his/her *IQ* by 1.5–2 *IQ* points by adding 10 weeks of Breastfeeding, depending on the values of other covariates.

Fig. 5 plots the conditional standard deviation of *IQ* against the conditional expectation for *IQ*, for each individual vector of covariates in the NLSY data

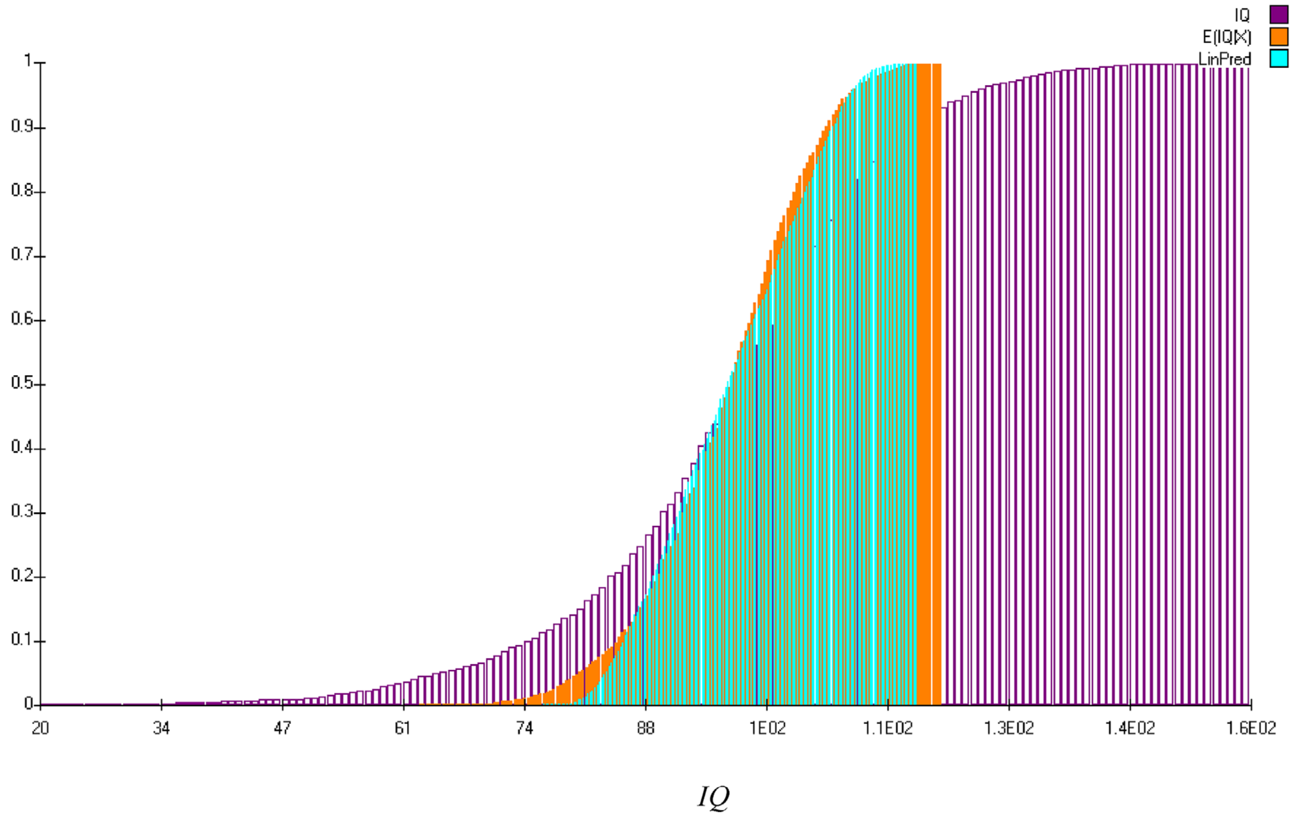

Fig 3. CDFs of IQ, the linear predictor of IQ, and $E(IQ\,|X\,)$.

Table V. Effect of breastfeeding duration on IQ for additional weeks δ

Effect of BFW on IQ: E_x [$E(IQ|X\backslash BF + \delta) - E(IQ|X)$] (32,000 samples)						
$\delta = 1$	3	5	10	15	20	25
0.303	0.456	0.636	0.895	1.122	1.316	1.554
Effect per week						
0.303	0.152	0.127	0.090	0.075	0.066	0.062
Effect per week (BFW > 1)						
0.192	0.104	0.096	0.072	0.063	0.056	0.054

set. The conditional standard deviation is evidently not constant. These standard deviations from the regression estimate are comparable to error distributions in ordinary linear regression and are often assumed to be constant even when the distributions are not normal. Fig. 5 shows that this assumption may be violated even when the copula is Gaussian. The non Gaussian behavior is solely attributed to non-Gaussian univariate margins.

4. EPICYCLES OF REGRESSION AND GROUND TRUTH

Epicycles were used in the Ptolemaic planetary model as ad hoc model adjustments to keep the Earth at the center of the universe while "saving the phenomena." The term is currently used to denote modeling tweaks not based on direct observations but motivated by other means.

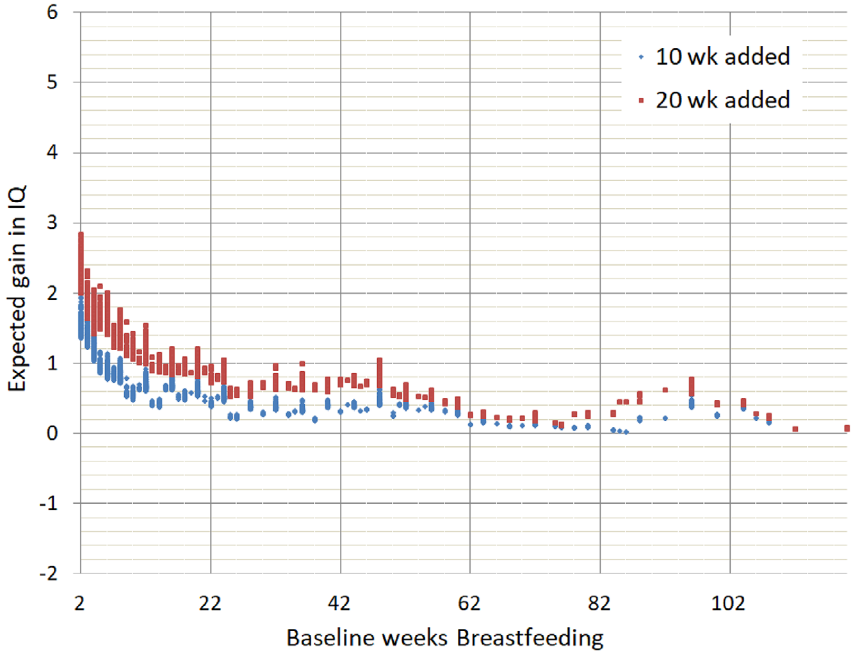

Fig 4. Scatter plot of $E(IQ \,|\, X \setminus X_{\mathrm{BFW}} + 10) - E(IQ \,|\, X)$ (blue) and $E(IQ \,|\, X \setminus X_{\mathrm{BFW}} + 20) - E(IQ \,|\, X)$ (red) against the baseline BFW starting at two weeks.

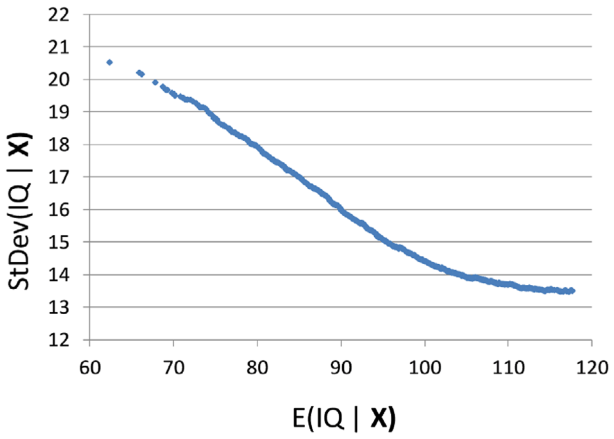

Fig 5. Conditional standard deviation versus conditional mean.

Traditional regression and vine regression follow different philosophies. In vine regression a vine copula density is chosen from a wide set of densities with arbitrary univariate margins and diverse dependence structures. Gaussian vines are the most convenient from the viewpoint of conditionalization but ignore features in the dependence structure like asymmetry and tail dependence. Based on heuristics, traditional regression selects an algebraic form for the regression function and estimates the coefficients from multivariate data. This section contrasts the two philosophies using the NLSY dataset. The goal is not to select an "optimal" functional form; absent a multivariate density this is a fool's errand. Nor is the intention to compile an exhaustive catalogue of regression epicycles; the set of possibilities is too large. Instead, we illustrate a few of the more obvious choices. From Fig. 2 we see that the standard deviation of *INC* is almost twice the mean. This might motivate a transformation to compress its values; a log or rank transformation could be considered. From Table I we see that *MAGE* and *CBIRTH* are not statistically significant, as their 90% confidence bands contain zero. This could motivate excluding these variables. Suspecting that the effect of weeks breastfeeding should taper off, one might add a quadratic term in *BFW*. Table II shows relatively high correlations of *MAFQT* with *INC*, *MGRADE*, and *BFW*, which could be an argument for including these three interaction terms. For good measure we add both the interaction terms and the quadratic term for *BFW*. Table VII compares the results.

In the presence of interactions and higher order terms, we cannot simply compare the coefficient of *BFW*. Instead, Table VII reports the average boost in predicted *IQ* resulting from giving each subject in the data 10 extra weeks of breastfeeding. The ratio of largest to smallest boost is 1.4; the effect of the choice of epicycles is in the order 40%. The most common metrics for judging model adequacy are fraction of explained variance (adjusted R^2) and residual standard deviation, each adjusted for the number of estimated parameters in the model. They give very little

Table VI. Effect of adding 10 weeks breastfeeding to baseline for Gaussian vine regression and linear regression, monetized according to (Grosse et al 2002) and (Gould 2009). For linear regression, the expected IQ points added to the population are 10×0.05×(Nr children at a given baseline value for weeks breastfeeding)

		Expected effects of 10 weeks additional breastfeeding						
	Weeks breastfed	**1**	**2**	**3**	**5**	**10**	**15**	**20**
	No. of children	203	181	126	58	59	9	143
Gaussian vine	Additional IQ points	714	287	178	56	33	6	89
	Monetary value [USD]	$12.7M	$5.1M	$3.2M	$1.0M	$0.59M	$0.11M	$1.6M
Linear model	Additional IQ points	104	93	64	30	30	5	73
	Monetary value [USD]	$1.8M	$1.6M	$1.1M	$0.53M	$0.54M	$0.08M	$1.3M

Table VII. The 7 multiple regressions from left to right are as follows: (a) predictors CBIRTH, BFW, MAGE, MGRADE, MAFQT, ln(INC); (b) similar to (a) with INC replacing ln(INC); (c) similar to (a) with rank(INC) replacing ln(INC); (d) MAGE and CBIRTH deleted; (e) model in (a) with addition of BFW^2; (f) model in (a) with addition of three interaction terms MAFQT with ln(INC), MGRADE, BFW; (g) model in (a) with the three interactions and BFW^2

	Transform INC			Exclude	Higher order	Interactions	
Ave IQ boost BFW = 10	ln(INC) (a)	INC (b)	Rnk(INC (c))	CBIRTH,MAGE (d)	BFW^2 (e)	MAFQT(Ln(INC)+ MGRADE+BFW) (f)	+BFW^2 (g)
	0.529	0.511	0.547	0.480	0.666	0.515	0.643
Adj R^2	0.239	0.234	0.242	0.238	0.239	0.241	0.241
Residual SD	15.477	15.523	15.442	15.492	15.476	15.452	15.451

Table VIII. Same multiple regressions as in Table VII, but applied to the proxy data set. The ground truth is added in the rightmost column

	Transform INC			Exclude	Higher order	Interactions		
Ave IQ boost BFW = 10	ln(INC) (a)	INC (b)	Rnk(INC) (c)	CBIRTH, MAGE (d)	BFW^2 (e)	MAFQT(Ln(INC)+ MGRADE+BFW) (f)	+BFW^2 (g)	Ground truth
	0.422	0.416	0.422	0.408	0.573	0.469	0.575	0.809
Adj R^2	0.244	0.241	0.244	0.244	0.244	0.245	0.245	
Residual SD	15.291	15.325	15.297	15.294	15.288	15.287	15.287	

guidance in choosing a model. It is difficult to imagine how the discussion over epicycles would ever conclude. The discussion of Ptolemy's epicycles went on for some 1,500 years.

What is needed is the ground truth. If the ground truth is the actual effect of giving each subject in this data set an additional 10 weeks breastfeeding, then the epicycle discussion runs no risk of ever concluding. However, the question of ground truth can be approached differently. Suppose we construct a proxy data set resembling the NLSY data, for which the actual density is known. We can easily generate such a proxy by drawing 3,179 samples from the density represented in Fig. 2. These samples have the same margins and similar dependence structure (see Table III, IV, 6), but they are NOT sampled from the distribution which generated the NLSY data. That distribution will forever be unknown to us. Rather, they constitute a similar multivariate data set for which the density is known. Therefore, the conditional expectations can be computed subject only to sampling error. We apply the epicycles of Table VII to this proxy data set and compare the results with the ground truth. The results are in Table VIII.

Adjusting the values in Table VIII for $BFW = 1$ we can infer that the coefficient of BFW in the case "$Ln(INC)$" in a linear regression is 0.042 and not 0.051 as in Table I. The value 0.042 lies well within the 90% confidence band for BFW in Table I (0.028, 0.077), but it is based on a separate sample from a

Table IX. Comparison of Gaussian C-vine and good fitting R-vine. For these comparisons the predictions were based on conditional medians rather than conditional means

	AIC	RMSE	MAD	Effect of BFW on IQ (δ = 10) (based on median)
Good fitting R-Vine	−10,619	15.664	11.65	0.062
Bayes net	−10,224	15.661	11.66	0.068

different distribution. The ratio between the largest and smallest effect of 10 additional weeks breastfeeding is again 1.4. The values for *adj* R^2 are a bit higher and those of standard error are a bit lower than in Table VII, reflecting the fact that the distribution in Fig. 2 is a bit smoother than the NSLY distribution and hence a bit more predictable. The ground truth effect in Table VIII, 0.809, is a bit lower than the comparable value 0.895 in Table V, V, which is based on the actual NSLY data.

The message from Table VIII is that none of the epicycle values are close to the ground truth, and the common heuristics of model fit are not pointing us in the right direction. With Table VII we have no way of knowing the truth and the epicycle discussion can continue unhindered. In Table VIII we know the ground truth and we can conclude with finality that the epicycle values are all wrong. That discussion has ended.

One feature of a preselected functional form for regression functions is that predictions can easily be made for covariate values which are out of sample, although one can debate whether this is a boon or a bane. For vine regression, regression functions are calculated for each individual in the data set. Predicting out of sample would require fitting a multivariate function to the regression functions. The difficulty in doing this is perhaps commensurate with the difficulty of predicting out of sample.

5. ASSESSING MODEL ADEQUACY

Assessing model adequacy in the case of Bayes nets is not straightforward. Indeed, such nets apply "Gaussian smoothing" of the dependence structure while preserving the one-dimensional distributions. Such models do not aspire to fit the data perfectly but rather to capture the molar features of the data. One way to assess adequacy in this sense is to fit a richer model in which key assumptions are relaxed and determine if the results change significantly.

The class of R-vines is the obvious candidate for this purpose. They relax the model structure shown in Fig. 2 and relax the restriction to Gaussian copulas. Recall that with the simplifying assumption different vine structures are not equivalent; some simplified vines will fit the data better than others. A good fitting R-vine for the NLSY data set was given in Cooke, Joe and Chang (2020). Suffice to say that only eight of the 21 bivariate copulas in the good fitting *R*-vine are Gaussian, and many of the others have asymmetry and tail dependence. The Akaike information criterion (*AIC*) value for the good fitting R-vine copula is -1.06×10^4, which is smaller than the *AIC* value -1.02×10^4 for the Gaussian copula, indicating a better fit (see Cooke, Joe & Chang 2020 for a fuller discussion)

The results for a good fitting R-vine and the Bayes net are shown in Table IX. The difference between the effect of breastfeeding duration on *IQ* based on the conditional means and conditional medians reflects the fact that the disproportionately large gains at the low end of *IQ* and breastfeeding duration are not captured by the difference: Median $(IQ|X\backslash BFW + 10)$ – Median$(IQ|X)$.

Even though the good fitting R-vine produces a better fit to the data than the Bayes net, for predicting *IQ* based on covariates, the R-vine has scarcely lower MAD and slightly higher MSE than the Bayes net. The value of the R-vine comparison in this case is to confirm the supposition that the Gaussian smoothing does a reasonable job in capturing the dependence between *IQ* and the covariates.

6. CONCLUSIONS

Vines can be a useful tool in regression analyses in two ways. First, they provide flexible and tractable classes of high-dimensional densities for representing multivariate continuous data. This may be done by Gaussian smoothing, which captures overall dependence while blurring out such features as asymmetry and tail dependence in the copula. Alternatively, a good fitting R-vine density can be fit to the multivariate data. Once a density is chosen, regression functions can be computed and the result of a policy change for a set of covariates can be readily computed. All regression models which are linear in the covariates will predict an effect that is linear in the covariates. Hence, breastfeeding for 25 years would increase the expected *IQ* by 65 points. Any "saturation effect" must be imposed from outside. In vine regression there is no agonizing over the epicycles of regression. The only question to be addressed is whether the density provides an adequate representation of the data. At present the best heuristic for answering this is to compare the results of a simple Gaussian smoothed BN, with a good fitting R-vine.

The second useful employment of vines in regression is to produce multivariate samples from a wide variety of multivariate densities which can serve to ground truth regression heuristics. From the example analyzed here, it appears that neither adjusted R^2 nor root mean square error provide reliable guides for finding the ground truth.

The NLSY is an observational study and the Bayes net model makes use of the univariate distributions of the variables as well as a smoothed representation of the dependence structure. Classical regression does not use this information and hence does worse for prediction.

Based on the Bayes net, the expected effects of breastfeeding on *IQ* depend on *IQ*, on the baseline level of breastfeeding, on the duration of additional breastfeeding and on the values of other covariates. A child given two weeks breastfeeding can expect to increase his/her *IQ* by 1.5–2 I*Q* points by adding 10 weeks of breastfeeding, depending on the values of other covariates. Such differentiated predictions cannot be obtained by regression models that are linear in the covariates.

ACKNOWLEGEMENT

The authors would gratefully acknowledge the detailed attention of the three anonymous referees who caught several errors and made several improvements. Harry Joe was sponsored by NSERC, Canada Discovery Grant No. 8698.

REFERENCES

Aas, K., & Berg, D. (2009). Models for construction of multivariate dependence — A comparison study, *European Journal of Finance*, *15*, 639–659.

Aas, K., Czado, C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, *44*(2), 182–198.

Acar, E. F., Genest, C., & Neshlehova, J. (2012). Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, *110*, 74–90

Bedford, T. J., & Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, *32*, 245–268.

Bedford, T. J., & Cooke, R. M. (2002). Vines — a new graphical model for dependent random variables. *Annals of Statistics*, *30*(4), 1031–1068.

Brechmann, E. C., & Schepsmeier, U. (2013). Modeling dependence with C- and D-vine copulas: The R package CDVine. *Journal of Statistical Software*, *52*(3), 1–27.

Chang, B., & Joe, H. (2019). Prediction based on conditional distributions of vine copulas. *Computational Statistics and Data Analysis*, *139*, 45–63.

Chollete, L., Heinen, A., & Valdesogo, A. (2009). Modeling international financial returns with a multivariate regime switching copula, *Journal of Financial Econometrics*, *7*(4), 437–480.

Cooke, R. M. (1997). Markov and entropy properties of tree and vine dependent variables. In *Proceedings of the American Statistical Association Section of Bayesian Statistical Science*. Alexandria, VA: American Statistical Association.

Cooke, R. M., Kurowicka, D., & Wilson, K. (2015). Sampling, conditionalizing, counting, merging, searching regular-vines. *Journal of Multivariate Analysis*, *138*, 4–18.

Cooke, R. M., Joe, H., & Aas, K. (2010). Vines arise. In D. Kurowicka & H. Joe (Eds.), *Dependence modeling: Handbook on vine copulae* (pp. 43–84). Singapore: World Scientific. Retrieved from https://www.sciencedirect.com/science/article/pii/S0047259/15000366

Cooke, R. M., Joe, H., & Chang, B. (2015) Vine regression RFF-DP 15–52.

Cooke, R. M., Joe, H., & Chang, B. (2020) Vine copula regression for observational studies. *AStA Advances in Statistical Analysis*, *104*, 141–167. https://doi.org/10.1007/s10182-019-00353-5

Czado, C., Brechmann, E. C., & Gruber, L. (2013). Selection of vine copulas. *Copulae in Mathematical and Quantitative Finance Lecture Notes in Statistics*, *213*, 17–37.

Fischer, M., Köck, C., Schlüter, S., & Weigert, F. (2009). Multivariate copula models at work. *Quantitative Finance*, *9*(7), 839–854.

Grosse, S. D., Matte, T. D., Schwartz, J., & Jackson, R. J. (2002) Economic gains resulting from the reduction in children's exposure to lead in the United States. *Environmental Health Perspective*, *110*, 563–569.

Gould, E. (2009) Childhood lead poisoning: conservative estimates of the social and economic benefits of lead hazard control. *Environmental Health Perspective*, *117*, 1162–1167.

Hanea, A. M., Kurowicka, D., Cooke, R. M., & Ababei, D. A. (2010). Mining and visualising ordinal data with non-parametric continuous BBNs, *Computational Statistics and Data Analysis*, *54*, 668–687.

Hobaek Haff, I., Aas, K., & Frigessi, A. (2010). On the simplified pair-copula construction - simply useful or too simplistic? *Journal of Multivariate Analysis*, *101*, 1296–1310.

Horta, B. L., & Victora, C. G. (2013). *Long-term effects of breastfeeding a systematic review*. Geneva, Switzerland: World Health Organization.

Jaworski, P., Durante, F., & Härdle, W. K. (2013). *Copulae in mathematical and quantitative finance. Proceedings of the workshop held in Cracow*. Lecture Notes in Statistics 213. Berlin: Springer.

Joe, H., Li, H., & Nikoloulopoulos, A. K. (2010). Tail dependence functions and vine copulas. *Journal of Multivariate Analysis*, *101*, 252–270.

Joe, H. (1994). Multivariate extreme-value distributions with applications to environmental data. *Canadian Journal of Statistics/La Revue Canadienne de Statistique*, *22*(1), 47–64.

Joe, H. (2014). *Dependence modeling with copulas*. Boca Raton, FL: Chapman & Hall, CRC.

Kraus, D., & Czado, C. (2017). D-vine copula based quantile regression, *Computational Statistics and Data Analysis*, *110*, 1–18.

Kurowicka, D., & Cooke, R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling*. Chichester, UK: Wiley.

Kurowicka, D, & Joe H. (Eds.). (2010). *Dependence modeling: Handbook on vine copulae* (pp. 43–84). Singapore: World Scientific.

Low, R. K. Y., Alcock, J., Faff, R., & Brailsford, T. (2013). Canonical vine copulas in the context of modern portfolio management: Are they worth it? *Journal of Banking & Finance*, *37*, 3085–3099

Morales Napoles, O. (2009). Bayesian belief nets and vines in aviation safety and other applications (PhD Thesis, Department of Mathematics, Delft University of Technology Delft, The Netherlands).

Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., & Graeler, B. (2019). Vine copula: Statistical inference of vine copulas. R package version 2.3.

Parsa, R. A., & Klugman, S. A. (2011). Copula regression, *Casualty Actuarial Society*, *5*(1), 45–54.

Sala-I-Martin, X. (1997). I just ran two million regressions. *The American Economic Review*, *87*(2), 178–183.

Stoeber, J., Joe, H., & Czado, C. (2013). Simplified pair copula constructions, limitations and extensions. *Journal of Multivariate Analysis*, *119*, 101–118.