November 2015 ■ RFF DP 15-52

DISCUSSION PAPER

Vine Regression

Roger M. Cooke, Harry Joe, and Bo Chang

1616 P St. NW
Washington, DC 20036
202-328-5000 www.rff.org

RESOURCES
FOR THE FUTURE

Vine Regression

Roger M. Cooke, Harry Joe, and Bo Chang

Abstract

Regular vines or vine copula provide a rich class of multivariate densities with arbitrary one dimensional margins and Gaussian or non-Gaussian dependence structures. The density enables calculation of all conditional distributions, in particular, regression functions for any subset of variables conditional on any disjoint set of variables can be computed, either analytically or by simulation. Regular vines can be used to fit or smooth non-discrete multivariate data. The epicycles of regression—including/excluding covariates, interactions, higher order terms, multi collinearity, model fit, transformations, heteroscedasticity, bias, convergence, efficiency—are dispelled, and only the question of finding an adequate vine copula remains. This article illustrates vine regression with a data set from the National Longitudinal study of Youth relating breastfeeding to IQ. Based on the Gaussian C-Vine, 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 2 weeks breastfeeding can expect to increase his/her IQ by 1.5 to 2 IQ points by adding 10 weeks of Breastfeeding, depending on values of other covariates. Averaged over the NLSY data, 10 weeks additional breastfeeding yields an expected gain in IQ of 0.726 IQ points. Such differentiated predictions cannot be obtained by regression models which are linear in the covariates.

Key Words: regular vine, vine copula, copula, C-vine, Gaussian copula, multivariate regression, heteroscedasticity, regression heuristics, National Longitudinal study of Youth, Breastfeeding, IQ

Contents

Vine Regression

Roger M. Cooke, Harry Joe, and Bo Chang*

1. Introduction

A Regular Vine (R-Vine) or Vine Copula (Cooke 1997, Bedford and Cooke 2002) is a tool for constructing high dimensional distributions with dependence. One dimensional margins can be taken from data, the dependence structure is represented by sets of bivariate and conditional bivariate copulas. A Wikipedia page provides a good informal introduction (https://en.wikipedia.org/wiki/Vine_copula). For definitions and properties of vines, see (Kurowicka and Cooke 2006, Kurowicka and Joe 2011, Joe, 2014), for their historical origins see (Joe 1994, Cooke, Joe and Aas 2010). Vines are most actively employed in financial mathematics (Aas and Berg, 2009, Aas et al 2009, Chollete et al 2009, Czado et al 2009, 2013, Fischer et al 2009, Jaworski et al 2012, Low et al 2013). Software has been developed at the TU Munich (Brechmann and Schepsmeier, 2013, Schepsmeier et al 2014), TU Delft (Hanea et al 2010), and the University of British Columbia (Joe 2014).

The number of labeled regular vines on n variables is quite large (Morales 2009, Cooke et al 2015):

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

and any absolutely continuous distribution on n variables may be represented on any regular vine with density written as a product of one dimensional marginal densities and copula densities (Bedford and Cooke 2001):

$$f_{1,2,\ldots,n}(x_1,\ldots,x_n) \quad = \quad f_1(x_1)\ldots f_n(x_n)\prod_{e\in\mathcal{V}} c_{e_1,e_2;D(e)} \tag{2}$$

where edge *e* in edge set $\mathcal{V}$ has conditioning set *D(e)* and conditioned variables *e1*, *e2*. 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

* Roger M. Cooke, Resources for the Future and University of British Columbia Department of Statistics, cooke@rff.org; Harry Joe, University of British Columbia Department of Statistics; Bo Chang, University of British Columbia Department of Statistics.

copula density functions do not depend on the values of the conditioning variables is often invoked, resulting in "simplified R-vines" (Hobaek Haff et al. 2010, Acar et al 2012, Stoeber et al 2013) . 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 be more suitable than others for a given distribution. A Gaussian R-vine, where the (conditional) copula are Gaussian, is always simplified, and can represent any absolutely continuous distribution with a multivariate Gaussian copula.

This article explores two uses of R-vines for regression. (A) R-vine models may be used to fit or to smooth data. 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. Given an R-vine density, the well-known epicycles of regression modeling, to wit: including/excluding covariates, interactions, higher order terms, multi collinearity, model fit, transformations, heteroscedasticity, bias, convergence, efficiency, simply do not arise (see also Sala-I-Martin 1997). (B) Because of expression (2), samples may be easily drawn from the vine distribution on *n* variables. Using one of the *2n-1* "implied sampling orders" (Cooke et al 2015), the sampling may be conditioned on values of initial segments of implied sampling orders, without extra computational burden. By conditioning on *n-k* variables, we may sample (or sometimes even compute) the joint regression function of the last *k* variables. Up to numerical and sampling error, this gives us the ability to determine exact regression functions for a rich family of multivariate densities, and this in turn affords the possibility of ground truthing various regression heuristics.

This article illustrates both possibilities, and focuses on Gaussian vines. Gaussian vines are not intended to fit the data, and they miss features like tail dependence and asymmetry (Joe et al 2010). On the other hand they often do a reasonable job of representing rank correlations and enable fast conditionalization. More detail on vine regression is found in (Kraus and Czado, 2015, Parsa and Klugman 2011).

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 an optimal R-vine for this dataset, and section 6 concludes.

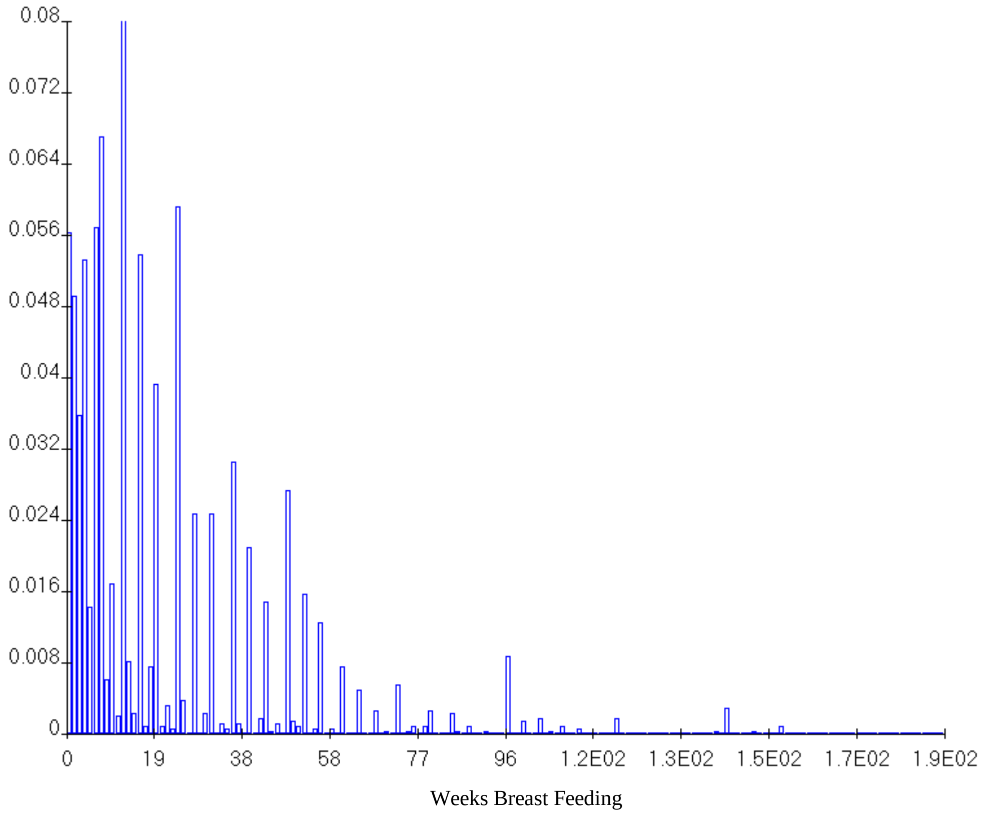
2. National Longitudinal Study of Youth (NLSY) data

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 and Victora 2013). The National Logitudinal Study of Youth (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 dataset of 2921 samples is obtained for child’s IQ, measured by the *Peabody Picture Vocabulary Test*, usually taken at age 8 to 10. The explanatory variables 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 (Mgrd), 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 breast feeding (Figure 1) range from 1 to 192. 2455 of the 2921 reported weeks breastfeeding are even, presumably a spurious mnemonic artifact. Among the odd numbers, only 59% are above 1 week. Many of the 1 week entries may indicate a failed attempt at breastfeeding, thereby conflating the effect of breast feeding duration with the effect of ever versus never breastfed. On the other hand, the effect of additional breast feeding is strongest for children with the smallest duration of breast feeding (see Figure 6). Hence, restricting to children with at least 2 weeks breast feeding probably leads to an under estimate of the effect of duration of breast feeding, whereas including children with less than 2 weeks breast feeding probably leads to an over estimate. In computing the effect of breast feeding duration on IQ (Tables 5,7) both options are given.

Figure 1. Histogram of Number of Weeks Breastfeeding in NLSY Data



The output of a simple linear regression is given below:

Table 1. Simple Linear Regression for NLSY Data

Regression Statistics	
Multiple R	0.49
R Square	0.24
Adjusted R Square	0.24
Standard Error	15.73
Observations	2921.00

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-263.73	251.04	-1.05	0.29	-755.97	228.51
CBIRTH	0.17	0.13	1.36	0.17	-0.08	0.43
BFW	0.05	0.01	3.77	0.00	0.02	0.08
MAGE	-0.26	0.14	-1.85	0.06	-0.53	0.01
MGRADE	0.52	0.14	3.77	0.00	0.25	0.78
MAFQT	0.27	0.01	21.22	0.00	0.24	0.29
INC	0.00	0.00	1.26	0.21	0.00	0.00

The rank correlation matrix of the NLSY data displays substantial "multicollinarity among the independent variables".

Table 2. Rank Correlation Matrix for NLSY Data

Rank Correlation Matrix from NLSY Data

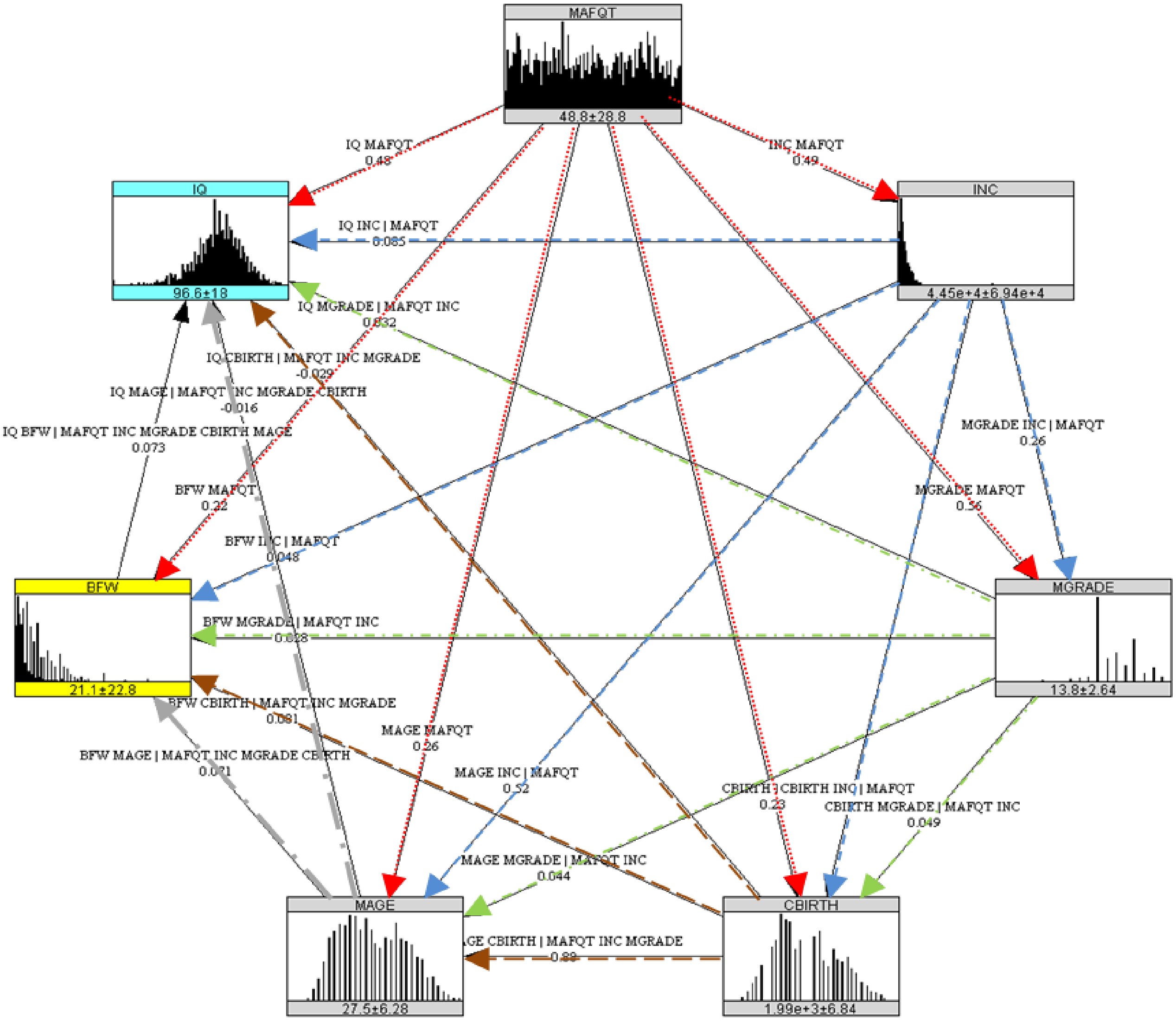
	Cbirth	BFW	Inc	Mafqt	Mage	Mgrade	IQ
Cbirth	1.00	0.17	0.63	0.24	0.93	0.30	0.16
BFW	0.17	1.00	0.17	0.25	0.19	0.18	0.18
Inc	0.63	0.17	1.00	0.50	0.63	0.49	0.33
Mafqt	0.24	0.25	0.50	1.00	0.28	0.56	0.50
Mage	0.93	0.19	0.63	0.28	1.00	0.32	0.17
Mgrade	0.30	0.18	0.49	0.56	0.32	1.00	0.32
IQ	0.16	0.18	0.33	0.50	0.17	0.32	1.00

3. Vine Regression

A C-vine is constructed using the one dimensional empirical distributions (ordinal data mining the original data in UNINET) with IQ in the top node. The C vine pictured as a saturated continuous non-parametric Bayesian Belief Net (CNPBBN) is shown in Figure 2. For details on CNPBBNs and their relation to R-vines, see (Kurowicka and Cooke 2006). Suffice to say that this C-vine is rooted at Mafqt. This is shown by the fact that it is the source node of all its arcs, and all rank correlations associated with these arcs are unconditional. The second root is Inc, all its arcs except that from Mafqt have Inc as their source, and these arcs are associated with conditional rank correlations given Mafqt. The third root is 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 and IQ is a sink for all arcs – its arcs all terminate at IQ.

If p_j denotes a partial rank correlation of variable j with sink node IQ in the conditioning set, $[1 - \Pi_j (1\text{-}p_j^2)]^{0.5}$ is the multiple rank correlation of IQ on all the other variables (Kurowicka and Cooke 2003, 2006a). Reading the values of partial rank correlations from Figure 1 yields $R^2 = 0.2423$, which is nearly equal to the R^2 in Table 1 *(0.2400)*.

Figure 2. C-vine for NLSY Data



Depicted as continuous non-parametric BBN, Partial rank correlations are shown. The C vine is built of nested trees, each with one node of maximal degree. The first tree is dotted red, the second is dashed blue, third, dot-dash green, fourth, long-dash brown, fifth, longdash-dot gray, sixth, black.

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 2. More precisely, we transform each variable X_i with CDF F_i to standard normal as $Z_i = \Phi^{-1}F_i(X_i)$, where Φ is the standard normal CDF. $\boldsymbol{Z} = (Z_1, \ldots Z_n)$ is not multivariate normal, but we consider a multivariate normal vector $\boldsymbol{Z'}$ with the same covariance matrix as $\boldsymbol{Z}$. Figure 2 shows partial rank correlations of $\boldsymbol{Z'}$; The rank correlation matrix of $\boldsymbol{Z'}$ is given in Table 3. It can be shown that the partials in Figure 2 uniquely determine the rank correlation matrix (Bedford and Cooke 2002). Together with the one dimensional margins and Gaussian copula with associated rank correlations assigned to each arc, these uniquely determine the joint distribution (Kurowicka and 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 $(X_1,\ldots X_n)$.

Table 3. Rank Correlation Matrix from Gaussian C-vine

Rank Correlation Matrix from C-vine							
	Cbirth	BFW	Inc	Mafqt	Mage	Mgrade	IQ
Cbirth	1.00	0.14	0.57	0.23	0.92	0.28	0.13
BFW	0.14	1.00	0.15	0.22	0.17	0.16	0.17
Inc	0.57	0.15	1.00	0.49	0.57	0.46	0.31
Mafqt	0.23	0.22	0.49	1.00	0.26	0.56	0.48
Mage	0.92	0.17	0.57	0.26	1.00	0.29	0.15
Mgrade	0.28	0.16	0.46	0.56	0.29	1.00	0.31
IQ	0.13	0.17	0.31	0.48	0.15	0.31	1.00

Not all correlation matrices are rank correlation matrices, and normal rank correlation matrices are sparse within the set of rank correlation matrices. (Lewandowski 2008, Lewandowski et al 2009, Joe 2006). Hence the differences between Tables 2 and 3 are not surprising and reflect the result of Gaussian smoothing.

Using the Gaussian copula, with any given any 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 2921 individuals in the data set, a Gaussian smoothed predictor of IQ is found, denoted *E(IQ|X)*. Table 4 compares the root mean square error and mean absolute deviation of the Gaussian smoothing prediction (*E(IQ|X)*) and the linear prediction with coefficients from Table 1, applied to the NLSY data.

Table 4. Root Mean Square Error and Mean Absolute Deviation of the Gaussian Smoothing Prediction (E(IQ|X)) and the Linear Prediction with Coefficients from Table 1, Applied to the NLSY Data

	E(IQ|X)	Linear Prediction
RMSE	15.64	15.71
MAD	11.70	11.77

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 income, etc. and this should influence our expectation of IQ.

One is often interested in a different question: "*If we change the BFW for an individual, how might that affect the 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, etc. 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, per week breast feeding, is then found by integrating this expected scaled difference over the population:

$$Effect\ BFW\ on\ IQ = E_{\boldsymbol{X}}(1/\delta)\ [\ E(IQ \mid \boldsymbol{X} \setminus X_{BFW}+\delta\,) - E(IQ \mid \boldsymbol{X})\ \ ]. \qquad (3)$$

In other words, we integrate the scaled difference of two regression functions which differ only in that one has δ weeks more breast feeding 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 in Gaussian C-vines.

Figure 3 shows the CDFs of IQ, $E(IQ \mid \boldsymbol{X})$ and also the CDF of the linear regression predictor of IQ from Table 1. $E(IQ \mid \boldsymbol{X})$ and the linear regression predictor are comparable, except on the low IQ end. The scatter plot in Figure 4 of both predictors against IQ confirms that $E(IQ \mid \boldsymbol{X})$ does a bit better at low IQ values.

Figure 3. CDF of IQ, the Linear Predictor of IQ, and E(IQ | X)

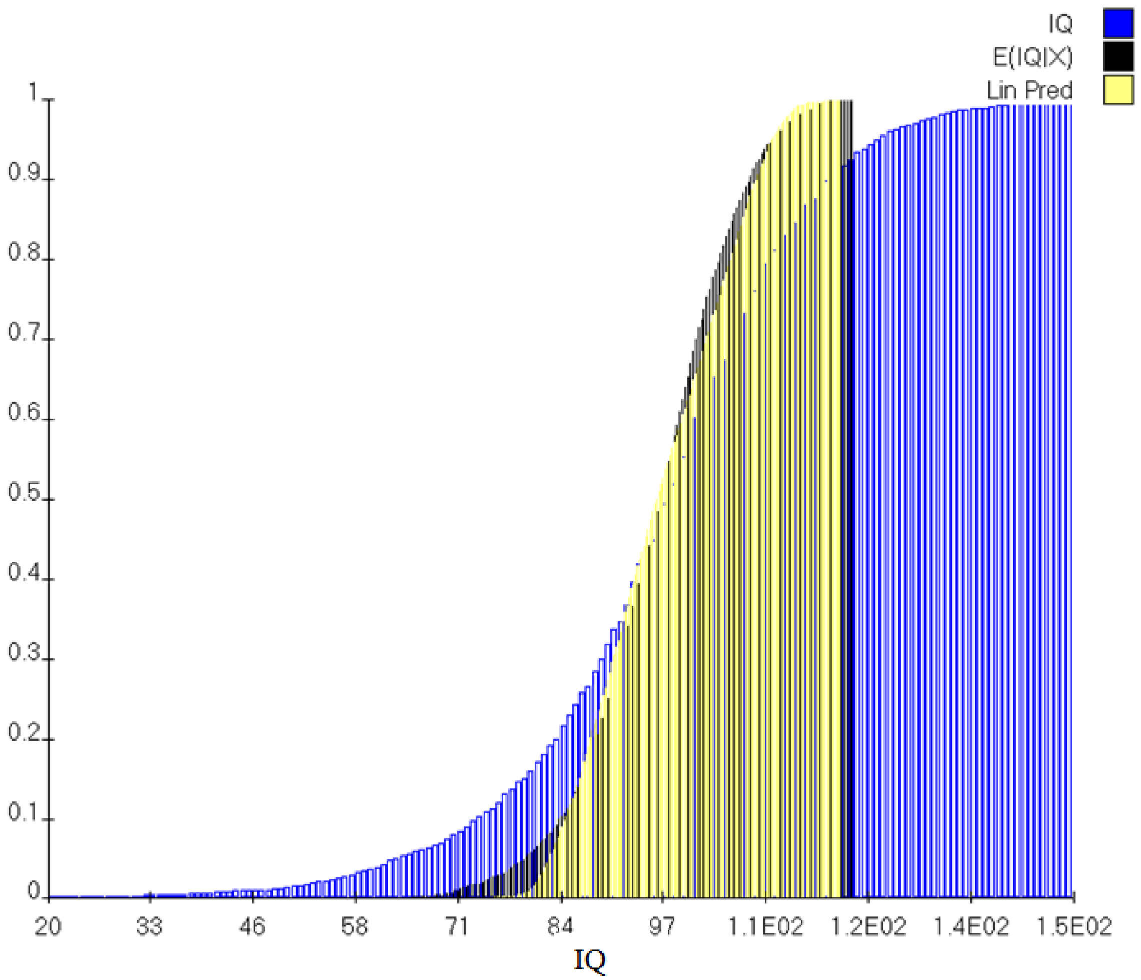
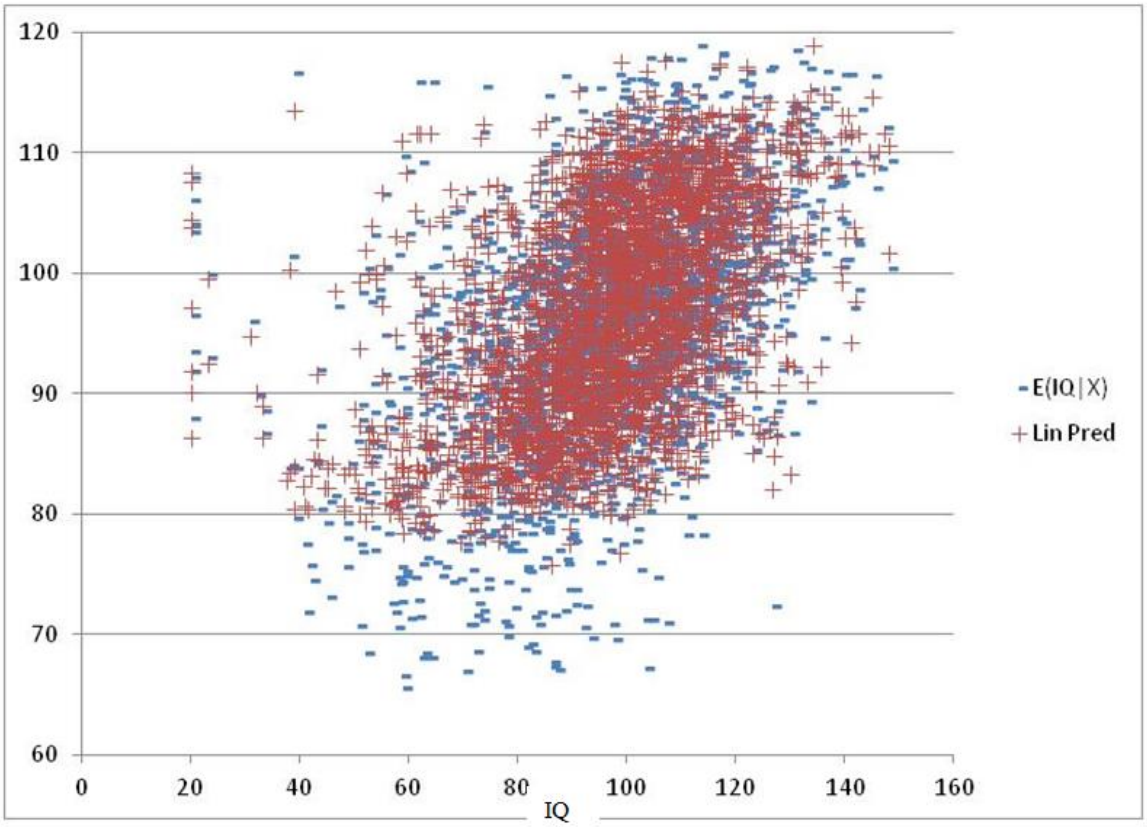


Figure 4. Scatter Plot of the Linear Predictor of IQ, and E(IQ | X) against IQ



The linear predictor assumes that the effect of breast feeding on IQ is linear; one additional week breastfeeding adds *0.05* IQ points, adding *25* years of breast feeding adds *65* IQ points. By varying δ in (3), vine regression avoids such implausible predictions. Table 5 shows the effect of breast feeding duration on IQ for values of δ from 1 to 25. To compare with the linear predictor, the effect also scaled per week added. Results of including and excluding individuals with *BFW*=1 are also shown.

Table 5. Effect of Breastfeeding Duration on IQ for Additional Weeks δ

Effect of BFW on IQ: E_x [E(IQ|X\BF+δ)-E(IQ|X)] (32,000 samples)						
δ = 1	3	5	10	15	20	25
0.2895	0.4533	0.6357	0.8947	1.1216	1.2984	1.5293
Effect Per Week						
0.2895	0.1511	0.1271	0.0895	0.0748	0.0649	0.0612
Effect Per Week (BFW > 1)						
0.1974	0.1071	0.0982	0.0726	0.0630	0.0554	0.0534

The differences of the regression functions *E(IQ |**X**)* and *E(IQ |**X**\BFW+10)* scattered against *BFW* (for *BFW>1*) show that the effect of additional weeks of breast feeding is greatest for low values of *BFW* (Figure 6). A child given 2 weeks breastfeeding can expect to increase his/her IQ by *1.5* to *2* IQ points by adding 10 weeks of Breastfeeding, depending on values of other covariates.

Figure 6. CDFs of E(IQ | X \ X_{BFW}+10) and E(IQ | X)

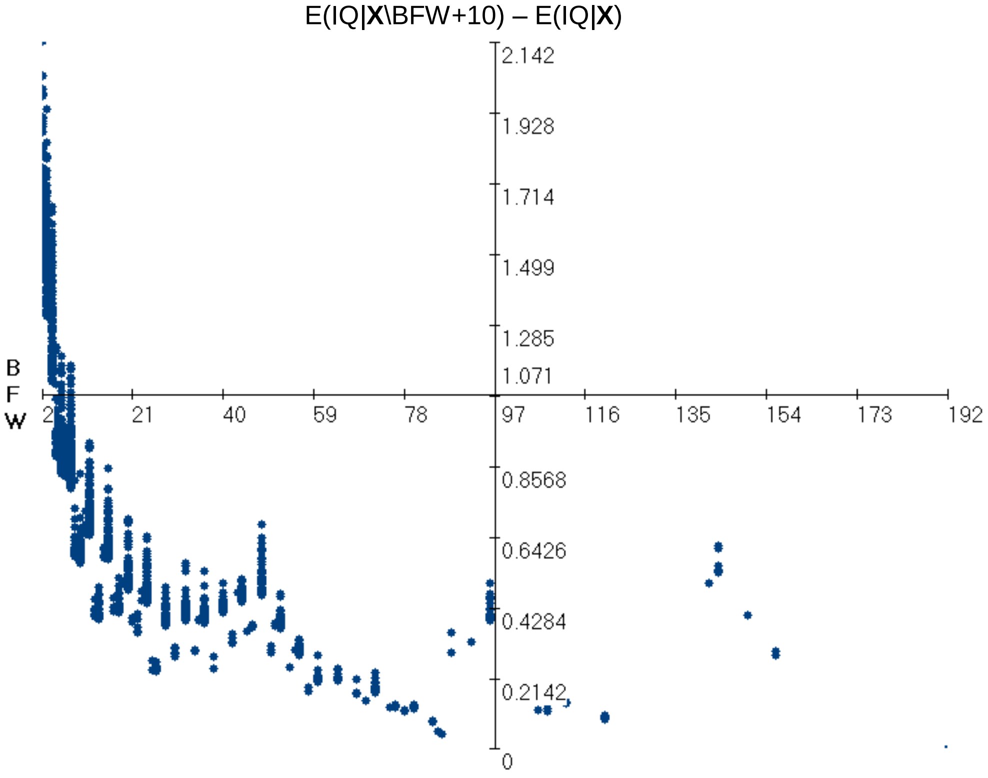
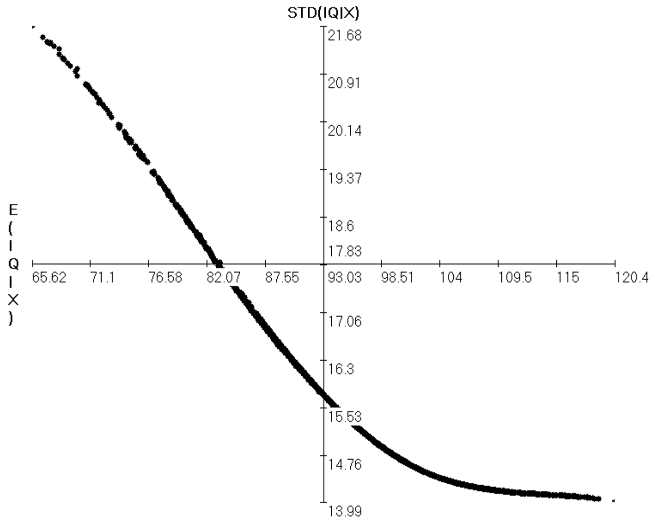


Figure 7 plots the conditional standard deviation of IQ against the conditional expectation for IQ, for each individual vector of covariates in the NLSY data set. The conditional standard deviation is evidently not constant.

Figure 7. Scatter Plot of Conditional Mean and Conditional Standard Deviation of IQ as Function of Covariate Values in NLSY Data



4. Ground Truth for Regression Heuristics

The above results are obtained with the actual NLSY data. We now replace the NLSY data with a data set sampled 2921 times using the Gaussian C-vine. This yields a data set with the same univariate margins, and with dependence similar to the NLSY data set for which the "ground truth density function" is known. Applied to this set, *E(IQ|X)* is the ground truth regression function, and can be estimated by drawing conditional samples - as many as desired. The results below are based on 1,000 conditional samples. Standard linear regression on the ground truth data yields the coefficients in Table 6, which resemble those in Table 1.

Table 6. Linear Regression on Ground Truth Data Set

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-301.70	209.57	-1.44	0.15	-712.63	109.22
CBIRTH	0.19	0.11	1.82	0.07	-0.02	0.40
BFW	0.05	0.01	3.95	0.00	0.03	0.08
MAGE	-0.23	0.12	-1.94	0.05	-0.46	0.00
MGRADE	0.23	0.13	1.69	0.09	-0.04	0.49
MAFQT	0.28	0.01	22.85	0.00	0.26	0.30
INC	0.00	0.00	1.98	0.05	0.00	0.00

Table 7 shows the results of computing the effect of BFW using the scaled differences of regression functions, analogous to eqn (3), but using regression coefficients like those in Table 6 for the linear model (this is 7 regressors counting the intercept) The model "*Linear + BFW^2*" simply adds *BFW^2* to the set of regressors. "2^{nd} order interactions" uses the regressors and all 2-way interactions *7 + 15 = 22* regressors). "4th Backwards" model starts with the 4th order model and eliminates regressors using R's backwards function to arrive at *34* regressors (R Core Team 2014). These include twelve pair wise, twelve 3-way and three 4-way interactions.

Table 7. Regressions on Data from Gaussian C-vine, and Scaled Expected Difference of Regression Functions, with and without adding 10 weeks to BFW (eqn 3) ("NormCop"). The linear model is form Table 6. The model "Linear + BFW^2" adds BFW^2 to the set of regressors. "2^{nd} order interactions" uses the regressors of Table 6 and all 2-way interactions, 7 + 15 = 22 regressors. "4th Backwards" model starts with the 4th order model and eliminates regressors using R's backwards function to arrive at 34 regressors.

C-Vine Normal Copula Regressions (δ=10)				
Model	**adjusted R^2**	**nr regressors**	**Effect BFW**	**Effect BFW, >1**
Linear	0.251	7	0.0518	0.0517
Linear+BFW^2	0.252	8	0.0885	0.0804
2^{nd} order interactions	0.256	22	0.0591	0.0545
4th Backwards	0.262	34	0.0381	0.0367
NormCop	**Ground Truth**	**6**	**0.0869**	**0.0868**

Note that the adjusted R squared values increase slightly with the number of regressors, indicating successively better fits, after accounting for the addition of regressors. The NormCopula samples the actual density used to generate the data, and thus (up to sampling fluctuations) constitutes the ground truth. Adding only the BFW^2 regressor comes closest to the ground truth, even though this model would not be selected by the adjusted R squared heuristic. Although the NormCop model has *6* regressors, it has more parameters if one were to count the parameters of the normal copula, though these parameters are stipulated by the C-vine generating the data and not estimated in this exercise.

Finally, Table 8 compares the mean square error and mean absolute deviation of the predictors from Table 6.

Table 8. Root Mean Square Error and Mean Absolute Deviation for Predictors in Table 6

	RMSE	**MAD**
Linear	15.46	11.82
Linear+BFW^2	15.43	11.80
2nd order interactions	15.42	11.90
4th Backwards	18.54	13.54
NormCop	15.30	11.71

The mean square errors are a bit lower than those in Table 4; the ground truth data is smoother and hence more predictable.

A number of observations emerge from this comparison. First, the heuristics RMSE, MAD and Adjusted R squared point in opposing directions (see 4th Backwards). None of the heuristics leads us closer to the ground truth. Finally, models which are close in terms of adjusted R squared, RMSE or MAD are not particularly close in the predicted effect of breast feeding duration.

5. Optimal R-vines

Instead of Gaussian smoothing, as performed above, we could alternatively fit an optimal R-vine to the data. The search algorithm for finding an optimal R-vine is based on the representation of an R-vine as a nested set of trees, in which the edges of the first tree are the nodes of the second tree, and so on. Edges belonging to the same tree are associated with conditional constraints of the same order. Referring to the C-vine in Figure 2, the first tree is associated with unconditional constraints (dotted red arrows). The second tree consists of constraints conditional on a single variable (in this case Mafqt, dashed blue arrows), and so on. The search proceeds in two steps:

1. Based on the correlation matrix of normal scores, which can be converted to different partial correlation vines, we look for an "optimal" vine that has large absolute correlations in tree 1, large absolute partial correlations in low order trees starting with tree 2 and small absolute partial correlation in higher-order trees subject to (BFW,IQ|other variables) being the edge of the last vine tree. This preliminary step assumes that we have bivariate Gaussian copulas on all edges of the vine.

2. Next, for the vine in step 1, we replace Gaussian copulas on edges of the vine with copulas that more appropriately account for tail asymmetry or tail dependence. With candidates such Frank, Gumbel, Gaussian, bb1 and MTJC and reflected versions of these (see Joe 1997)., we can find a good choice based on AIC for each edge of the vine (using the Munich VineCopula package for this step).

Conditionalizing R-vines is not yet well developed, but (Cooke et al 2015) introduce the notion of a "sampling order implied by an R-vine". Any implied sampling order may be sampled without extra computational burden, and the sampling may be conditionalized on any initial segment of the ordering, by simple plug-in conditionalization. These features can be used to support simple implementations of (3) for R-vines.

The optimal R-vine for the NLSY data based on steps 1 and 2 above is shown in Figure 8. This R-vine is not a C vine and does not decompose as a sequence of roots. Nonetheless, Mafqt is almost a first root, and Inc is a second root. It also has many non-Gaussian copula. The Akaike Information Criterion (AIC, Akaiki 1974) measures model fit, lower values indicating better fit. The AIC value for the optimal R-vine -10621 compares favorably to the value for the Gaussian vine, -10299. The R-vine results are computed with the U. British Columbia code which uses the conditional median instead of the conditional mean as predictor. The results for optimal R-vine and the Gaussian smoothed C vine are shown in Table 9. The difference between the effect of breast feeding 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 breast feeding duration are not captured by the difference *Median(IQ|X\BFW+10) – Median(IQ|X).*

Table 9. Comparison of Gaussian C-vine and Optimal R-vine.

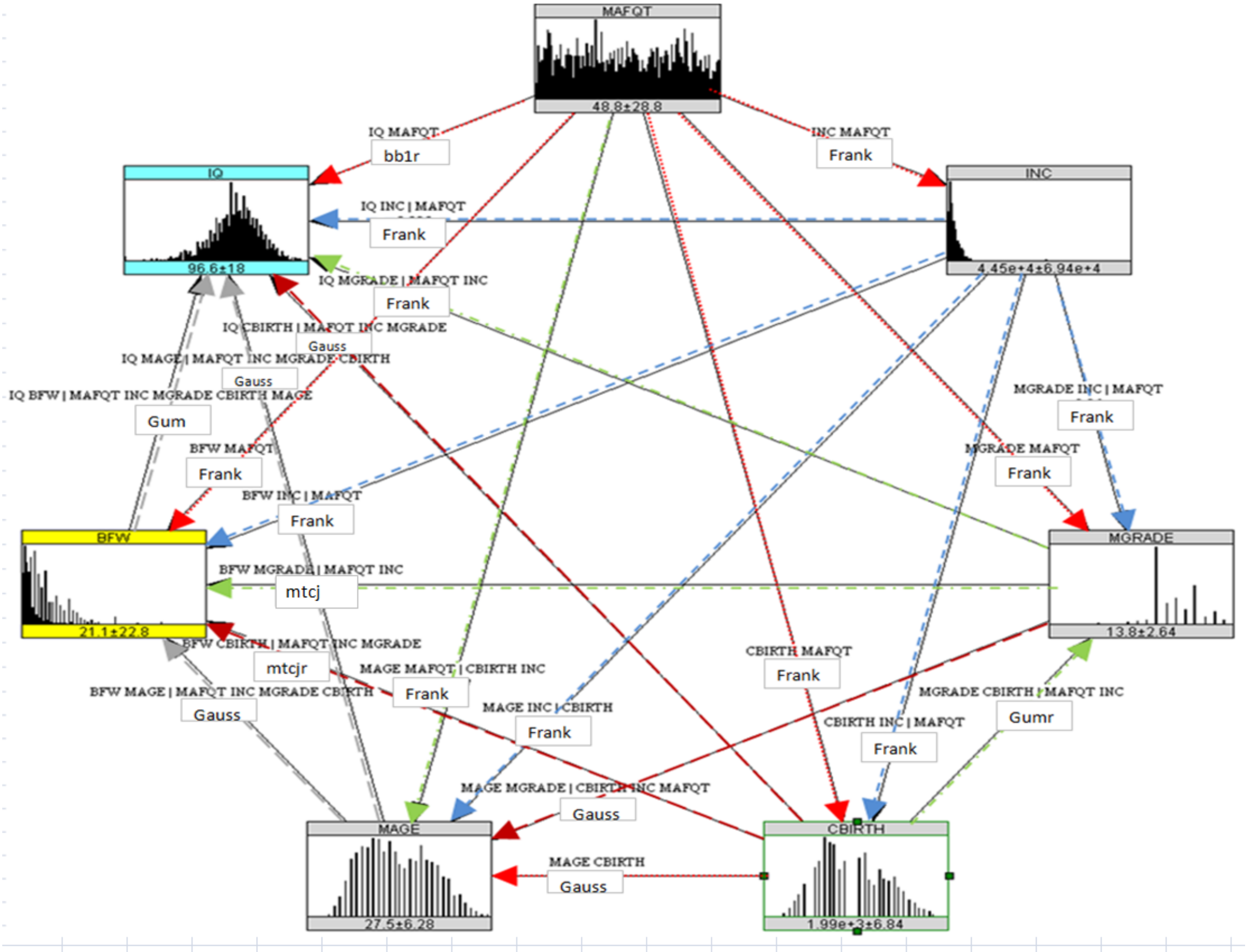
	RMSE	MAD	Effect of BFW on IQ ($\delta = 10$) (based on median)
Optimal R-Vine	15.66	11.65	0.062
Gaussian C-Vine	15.66	11.66	0.068

For these comparisons the predictions were based on conditional medians rather than conditional means.

Even though the optimal R-vine produces a better fit to the data than the Gaussian C vine, the R-vine has scarcely lower MAD than the Gaussian vine, and the RMSE's are effectively indistinguishable..

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.

Figure 8. Optimal R-vine for NLSY Data, Depicted as Continuous Non-parametric BBN with non-Gaussian copulae (mtcj = Mardia-Takahasi-Clayton-Cook-Johnson, Gum = Gumbel, suffix r = reflected or survival



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, an optimal 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, breast feeding for 25 years would increase 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 smoothed fit, with the best fitting R-vine. It would be useful to develop procedures for systematically checking robustness for different vine densities.

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 analysed here, it appears that neither adjusted R squared, nor mean square error nor mean absolute deviation provide reliable guides for finding the ground truth.

Based on the Gaussian C-Vine, 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 2 weeks breastfeeding can expect to increase his/her IQ by 1.5 to 2 IQ points by adding 10 weeks of Breastfeeding, depending on values of other covariates. Such differentiated predictions cannot be obtained by regression models which are linear in the covariates.

References

Aas K. and Berg D., (2009), Models for construction of multivariate dependence — a comparison study, *European J. Finance*, 15:639–659.

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

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

Akaike, H. *(1974),* "A new look at the statistical model identification" *(PDF), IEEE Transactions on Automatic Control 19 (6): 716–723,* doi*:*10.1109/TAC.1974.1100705, MR 0423716.

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

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

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

Chollete L., Heinen A. and Valdesogo A., (2009), Modeling international financial returns with a multivariate regime switching copula, Journal of Financial Econometrics, 2009, Vol. 7, No. 4, 437–480.

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

Cooke, R.M., Kurowicka, D. and Wilson, K. (2015) "Sampling, Conditionalizing, Counting, Merging, Searching Regular Vines" Journal of Multivariate Analysis doi:10.1016/j.jmva.2015.02.001 Available online online 14 February 2015, ISSN 0047-259X, http://dx.doi.org/10.1016/j.jmva.2015.02.001.

Cooke, Roger M., Joe, H. and Aas, K. (2010) Vines Arise, in Kurowicka and Joe (eds) Dependence Modeling: Handbook on Vine Copulae, World Scientific, Singapore, 978-981-4299-87-9, 981-4299-87-1, pp43-84. (http://www.sciencedirect.com/science/article/pii/S0047259X15000366)

Czado C., Min A., Baumann T. and Dakovic R., (2009), Pair-copula constructions for modeling exchange rate dependence. Technical report, Technische Universitaet Muenchen.

Czado, C., Brechmann, E.C., Gruber, L. (2013) Selection of Vine Copulas, Copulae in Mathematical and Quantitative Finance Lecture Notes in Statistics Volume 213, 2013, pp 17-37.

Fischer M., Kock C., Schluter S. and Weigert F., (2009), Multivariate copula models at work. Quantitative Finance, 9(7): 839–854.

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

Heinen A. and Valdesogo A., (2008), Canonical vine autoregressive model for large dimensions. Technical report.

Hobaek Haff,I., Aas, K. and 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, World Health Organization, ISBN 978 92 4 150530 7 (NL classification: WS 125)

Jaworski, P. Durante,F., Härdle, W.K., (2012) Copulae in Mathematical and Quantitative Finance: Proceedings of the workshop held in Cracaw, July 10-11, 2012, Lecture Notes in Statistics 213, Springer .

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

Joe H., (1997) Multivariate Models and Dependence Concepts. Chapman & Hall, London.

Joe H., (2006), Generating random correlation matrices based on partial correlations. J. of Multivariate Analysis, 97:2177–2189.

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

Joe, H (2014) Dependence modeling with Copulas, Chapman Hall, CRC, isbn 978-1-4665-8322-1

Kraus, Daniel, Czado, Claudia (2015) D-vine copula based quantile regression, Technische Universität München

Kurowicka and Joe (eds) (2011) Dependence Modeling: Handbook on Vine Copulae, World Scientific, Singapore, 978-981-4299-87-9, 981-4299-87-1, pp43-84.

Kurowicka D. and Cooke R.M., (2003), A parametrization of positive definite matrices in terms of partial correlation vines. *Linear Algebra and its Applications*, 372:225–251.

Kurowicka D. and Cooke R.M., (2004), Distribution-free continuous Bayesian belief nets. In *Mathematical Methods in Reliability*.

Kurowicka D. and Cooke R.M., (2006), *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley.

Kurowicka D. and Cooke R.M., (2006a), Completion problem with partial correlation vines. *Linear Algebra and Its Applications*, 418(1):188–200.

Kurowicka D. and Cooke R.M., (2007), Sampling algorithms for generating joint uniform distributions using the vine-copula method. *Computational Statistics and Data Analysis*, 51:2889–2906.

Kurowicka, D and Cooke, R.M.(2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley.

Lewandowski, D., (2008), *High Dimensional Dependence. Copulae, Sensitivity, Sampling*. PhD thesis, Delft Institute of Applied Mathematics, Delft University of Technology.

Lewandowski, D., Kurowicka D. and Joe H., (2009), Generating random correlation matrices based on vines and extended onion method, *J. Mult. Anal.*,100:1989–2001.

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 (2013) 3085–3099

Morales Napoles, Oswaldo (2009) *PhD Thesis Bayesian Belief Nets and Vines in Aviation Safety and Other Applications*. Department of Mathematics, Delft University of Technology, Delft, 2009.

Parsa, R.A., and Klugman, S.A. (2011) Copula Regression, Casualty Actuarial Society, Volume 5/Issue 1

R Core Team.(2014.) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria,

Sala-I-Martin, Xavier X. (1997) I Just Ran Two Million Regressions The American Economic Review Vol. 87, No. 2, Papers and Proceedings of the Hundred and Fourth Annual Meeting of the American Economic Association (May, 1997), pp. 178-183

Schepsmeier, U., Stoeber, J., Brechmann, E.C. and Graeler, B. (2014) Vine Copula:Statistical inference of vine copulas, R package version 1.3.

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

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